COSMO HUB on Hadoop

https://cosmohub.pic.es

A web portal to analyze and distribute cosmological data

J.Carretero, P.Tallada, J.Casals, M.Caubet, C.Neissner, N.Tonello, J.Delgado, F.Torradeflot, M.Delfino, S.Serrano, P.Fosalba

PIC port d'informació científica

ICE CSIC - IEEC

IFAE Institut de Física d'Altes Energies
25 years 1991-2016

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

IFAE - Pizza seminar - 25/01/2017

Outlook

- What is CosmoHub
- Motivation
- Drawbacks
- Hadoop solution
- Comparison
- Demo
- Conclusions & future work

COSMO HUB

Build your own Universe

Real-time data analysis of massive cosmological data without any SQL knowledge

Hundreds of millions of observed and simulated galaxies

Superfast queries means superfast results

Features to make you work faster and easier

Online plotting preview and data download

CosmoHub on PostgreSQL

- CosmoHub was first thought as a way to share data within two closed related projects, the Physics of the Accelerating Universe (PAU) and the Marenostrum Institut de Ciències de l'Espai simulations (MICE)
- It was built on top of a PostgreSQL relational database
- It was developed by people from the Institut de Ciències de l'Espai (ICE), the Port d'Informació Científica (PIC - www.pic.es), CIEMAT and IFAE
- It was hosted and operated at PIC

Old platform main features

- Custom catalogs without any SQL knowledge (CSV.BZ2 only)
- Plot & preview tool: small sample of data using a scatter plot or generate a 1D-histogram (query time limited to < 2’)
- Value-Added-Data ready to be downloaded

Some numbers

- CosmoHub is currently supporting four different cosmology projects:

MICE
Marenostrum Institut
de Ciències de l'Espai
Simulations

PAU

DARK ENERGY
SURVEY

euclid

- ~ 400 users
- ~ 1300 custom catalogs
- ~ 250 prebuilt downloads
- ~ 3 TiB hosted data
- > 10^9 objects

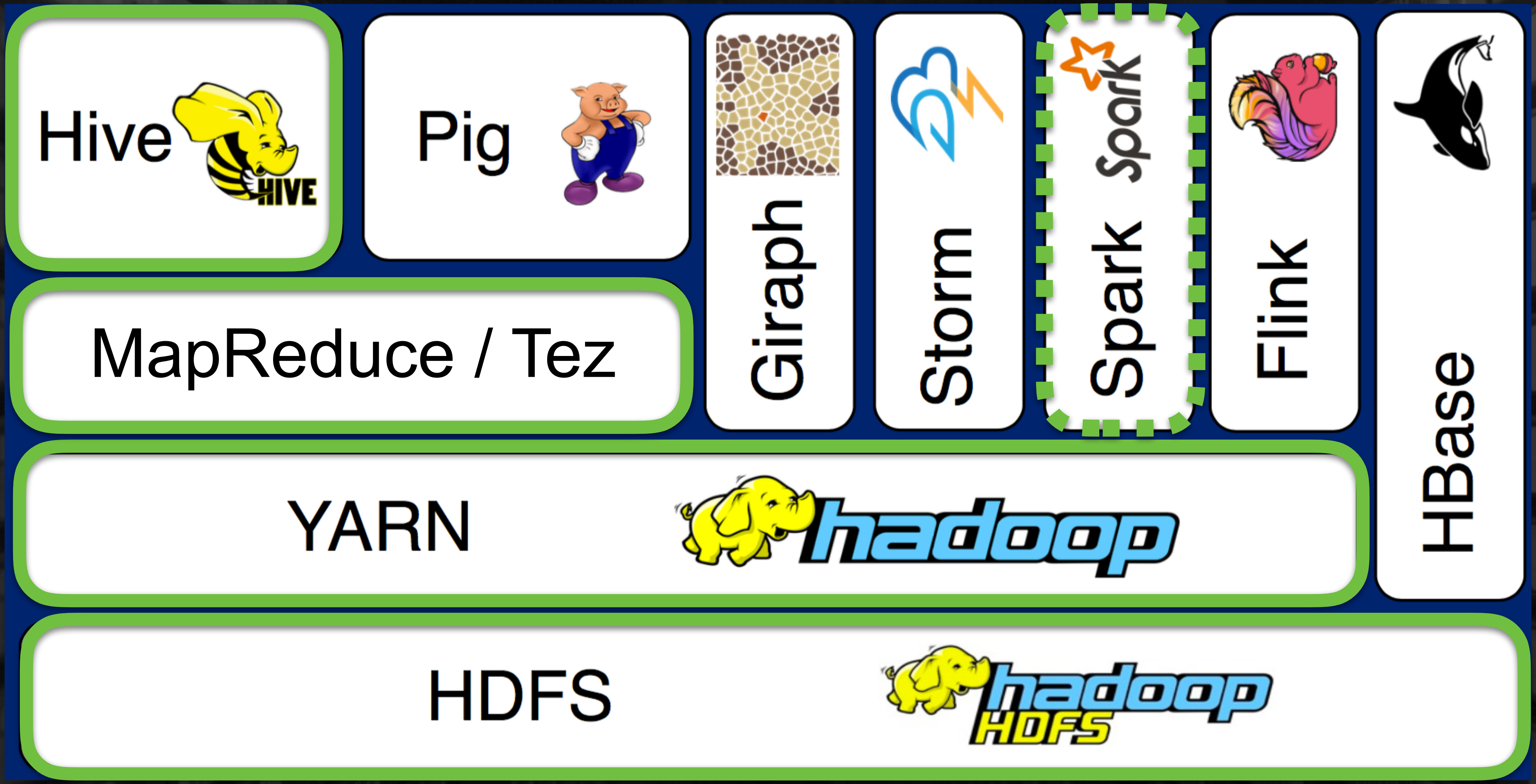
What happened?

- MICECATv2.0 catalog contains about 500M entries with more than 120 fields
- Managing large volume of data in PostgreSQL had some drawbacks:
  - Indices are not used for large datasets
    - Most of the queries lasted several hours
  - Changing the schema was very slow
  - Removing large subsets of data is very inefficient
- Future galaxy catalogs will contain a few 10^9 entries

Apache Hadoop & Hive

- Apache Hadoop:
  - one of the most popular Big Data platform
  - open-source software
  - based on commodity computer clusters
    - distributed storage and distributed processing
    - scalable from dozens up to even thousands of nodes
    - failure tolerance
- Apache Hive: query over massive data volumes

Hadoop stack



Hive

Pig

Giraph

Storm

Spark

Flink

HBase

MapReduce / Tez

YARN

HDFS

HORTONWORKS®

PostgreSQL vs Hadoop

- Larger relative gain in execution time for increasing complexity in datasets and/or as queries request larger data volumes

(Comparisons are odious. It is very likely to unjust to one or other of them)

Hadoop

- Nodes: 15
  - Cores: 12 (Intel Xeon X5650 @ 2.67 GHz) [180]
  - RAM: 24 GiB [360 GiB]
  - DISK: 1 TiB [15 TiB raw; ~5 TiB net]
  - Network: 1GbE

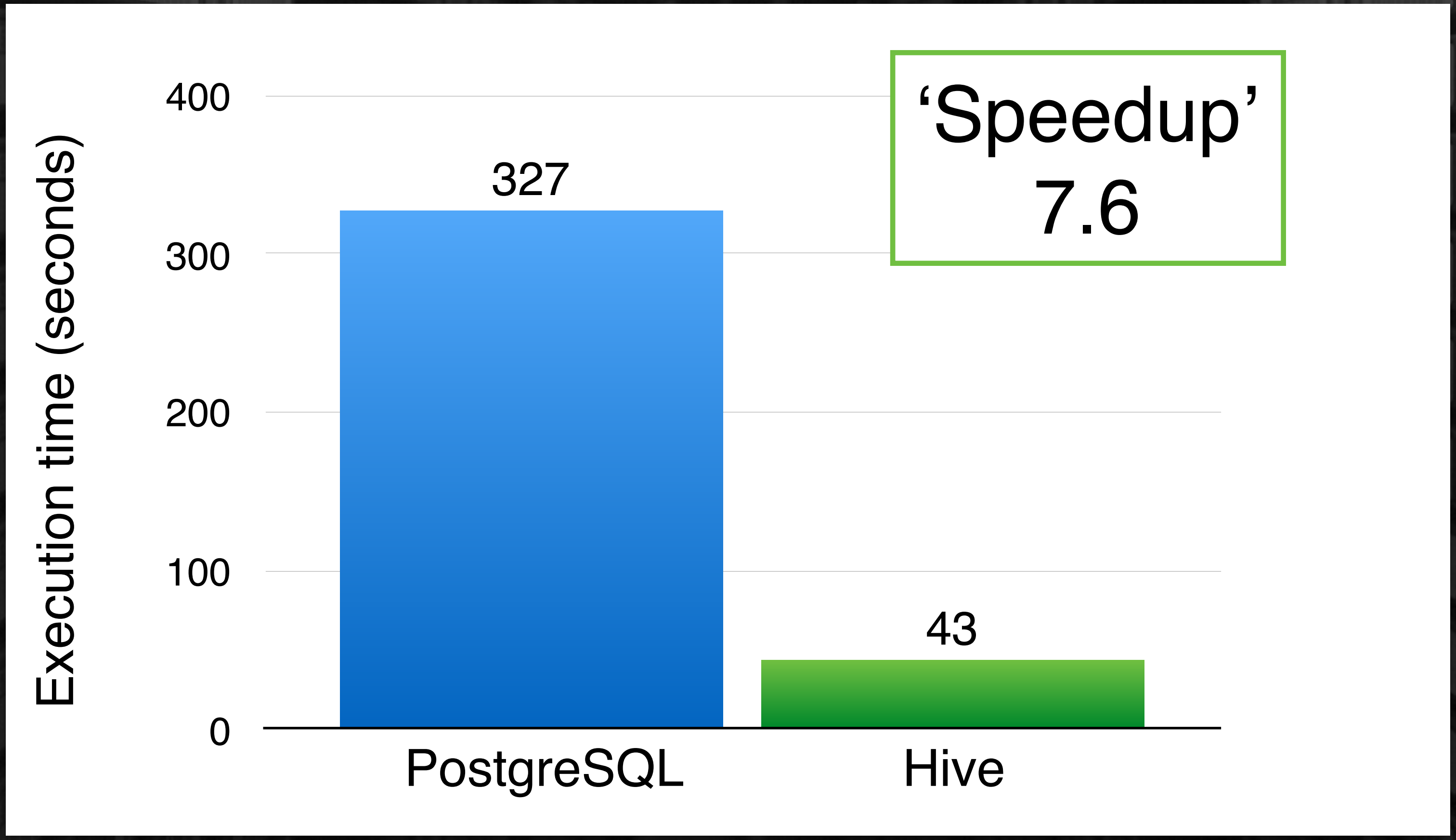
PostgreSQL

- Hardware:
  - Cores: 24 (Intel Xeon X5675 @ 3.07 GHz)
  - RAM: 96 GiB
  - DISK: 600 GB HDD x 8 (in RAID 6) ~ 3.6 TB net
  - Network 1GbE
- Software:
  - Scientific Linux 6.1
  - PostgreSQL 9.1

PostgreSQL vs Hadoop

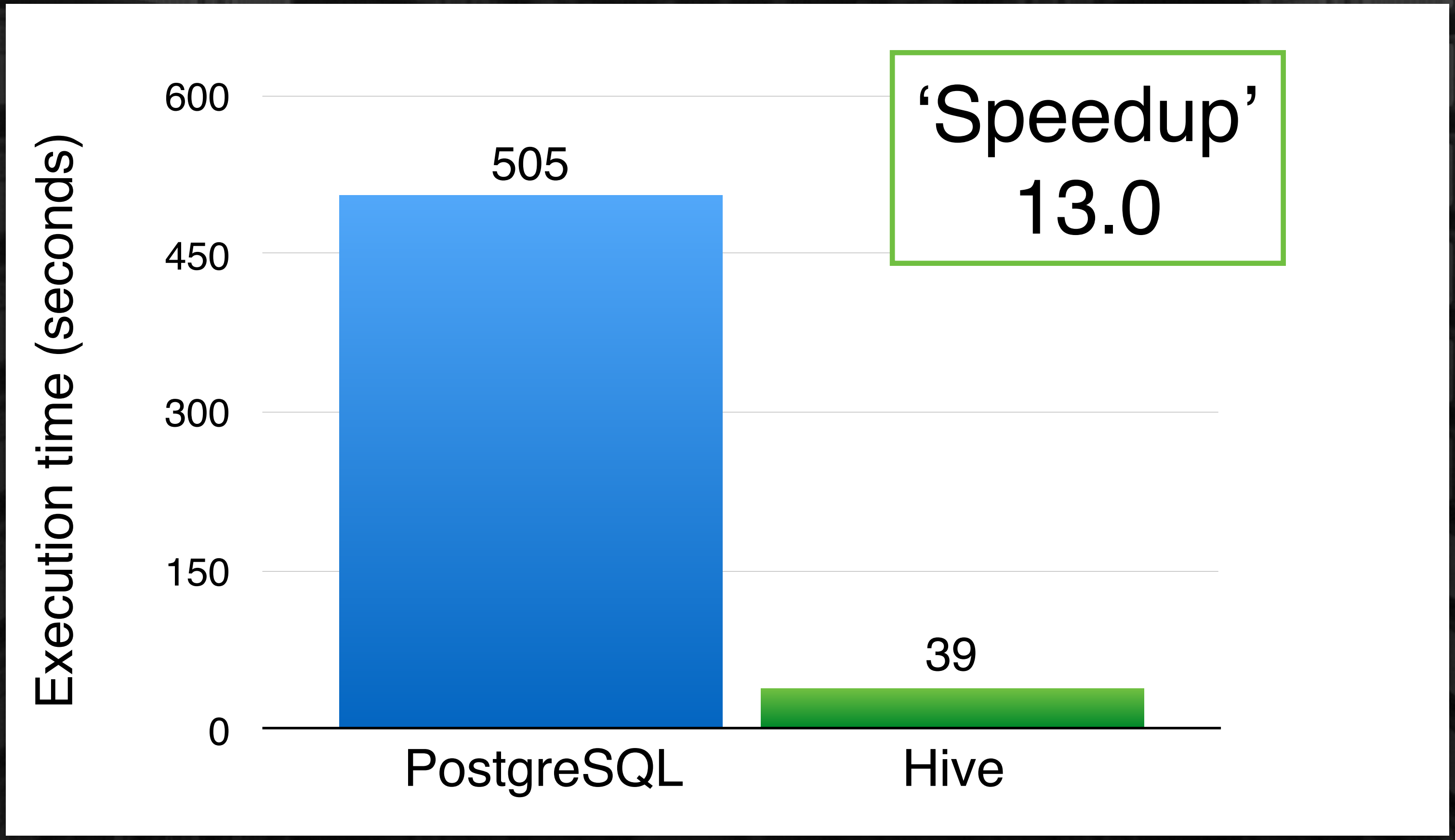
```
SELECT ra, dec, z, z_v, x_c, y_c, z_c FROM
micecatv1 WHERE x_c < 700 AND y_c < 700 AND
z_c < 700;
```

(~5.8M out of ~205M rows)



PostgreSQL vs Hadoop

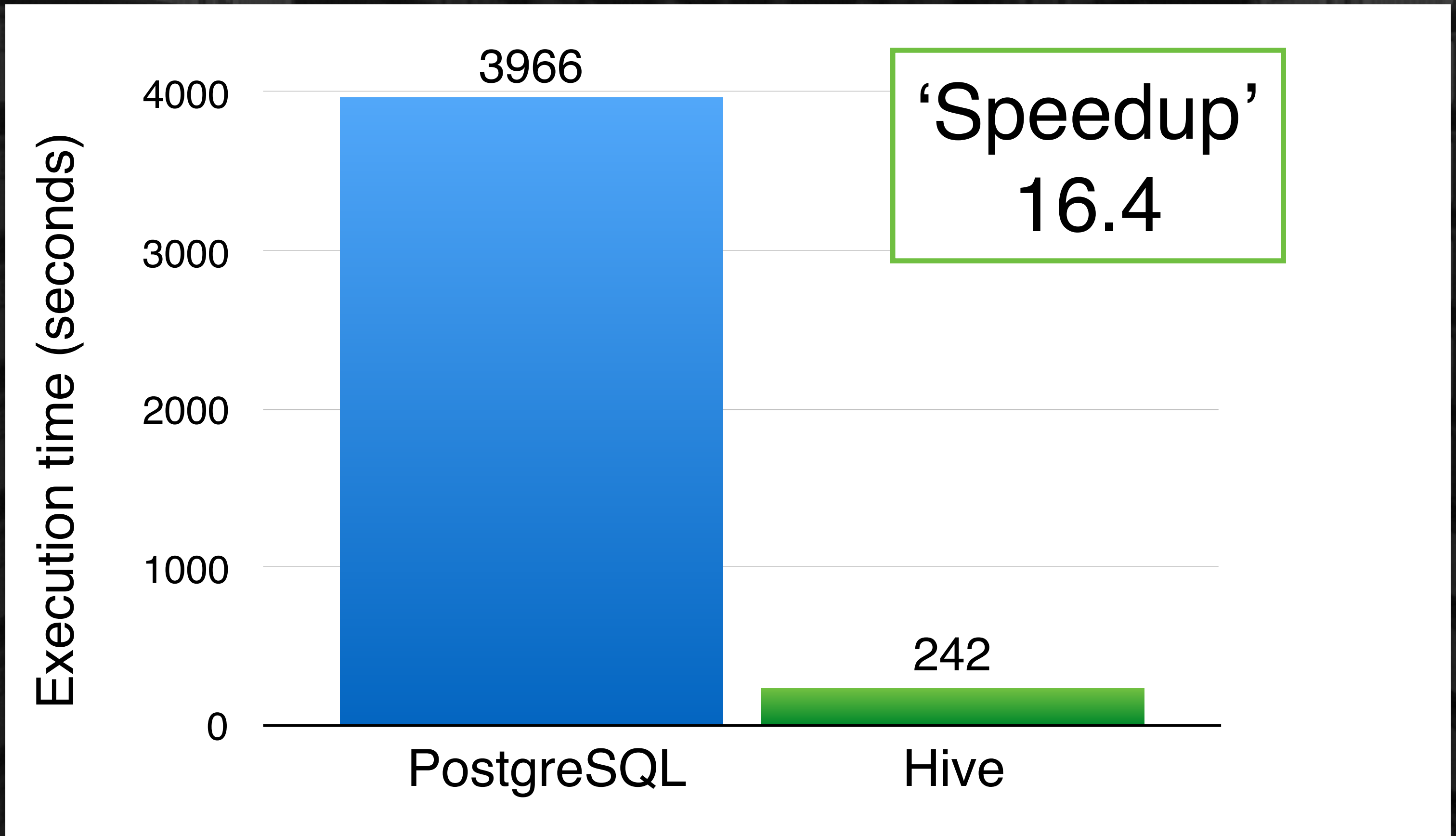
`SELECT x_c, y_c, z_c FROM micecatv1 WHERE x_c < 1e3 AND y_c < 1e3 AND z_c < 1e3;` ***(~16.5M out of ~205M rows)***



PostgreSQL vs Hadoop

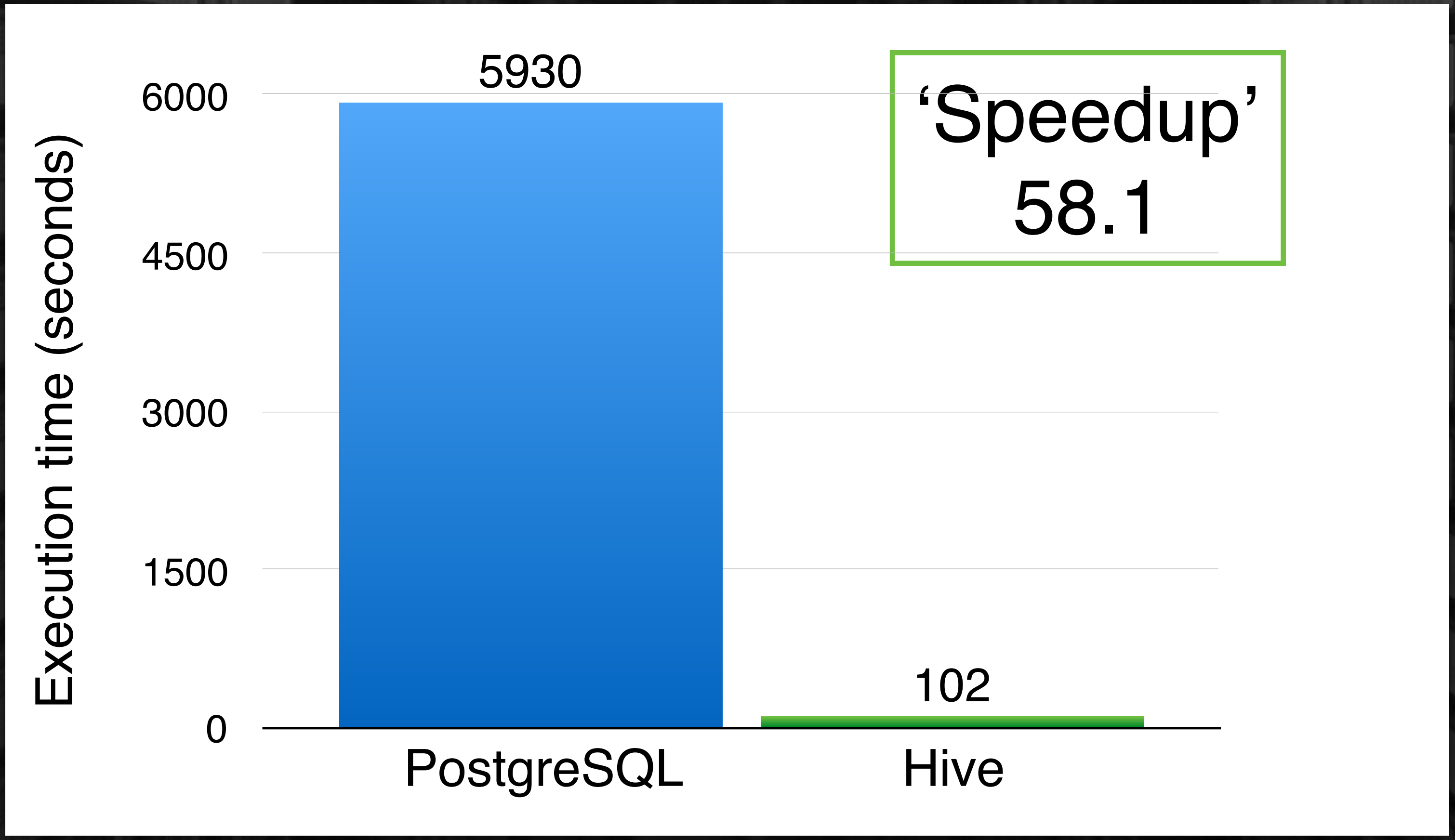
```
SELECT coadd_objects_id, ra, dec, mag_auto_i,
magerr_auto_i desdm_zp, mean_z_bpz, z_mc_bpz
FROM des_y1a1 WHERE modest_class = 1 AND
flags_gold = 0 AND flags_badregion = 0;
```

(~81.9M out of ~137M rows)



PostgreSQL vs Hadoop

```
SELECT ra_gal, dec_gal, kappa, gamma1, gamma2
FROM micecatv2 WHERE lmhalo >= 12.16 AND
flag_central = 0 AND z_cgal > 0.4 AND z_cgal
< 0.6; (~25.9M out of ~500M rows)
```

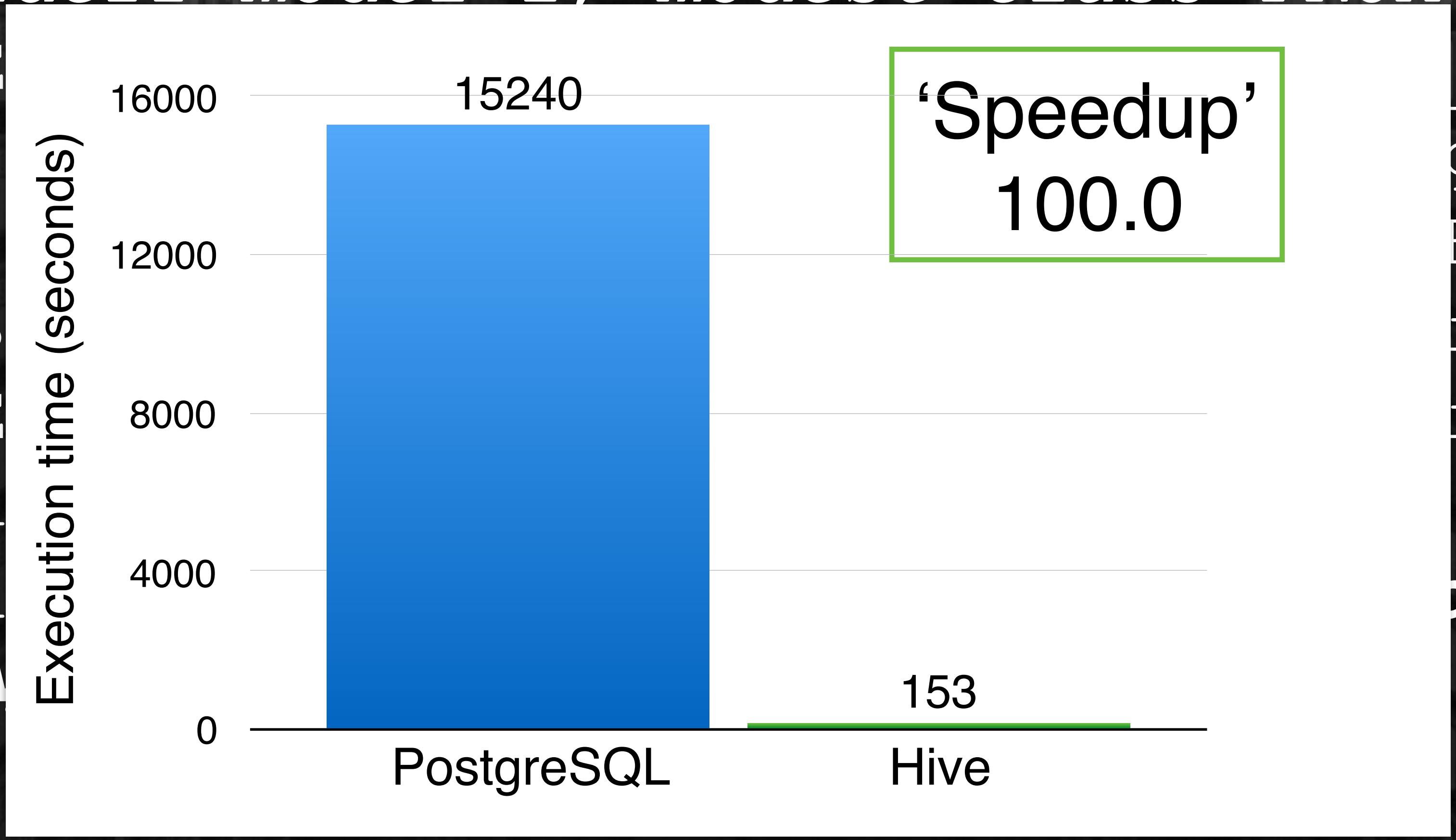


PostgreSQL vs Hadoop

```
SELECT coadd_objects_id, ra, dec, mag_auto_g,
mag_auto_r, mag_auto_i, mag_auto_z,
mean_z_bpz, mode_z_bpz, median_z_bpz,
z_mc_bpz, t_b, spread_model_i,
spreaderr_model_i, modest_class FROM des_y1a1
WHERE mag_auto_i > 17.5 AND mag_auto_i < 22
AND (flags_badregion <= 3 and flags_gold = 0)
AND ((mag_auto_g - mag_auto_r) BETWEEN -1.
and 3.) AND ((mag_auto_r - mag_auto_i)
BETWEEN -1. and 2.5) AND ((mag_auto_i -
mag_auto_z) BETWEEN -1. and 2.) AND (ra < 15
or ra > 290 or dec < -35); (~34.8M out of
~137M rows)
```

PostgreSQL vs Hadoop

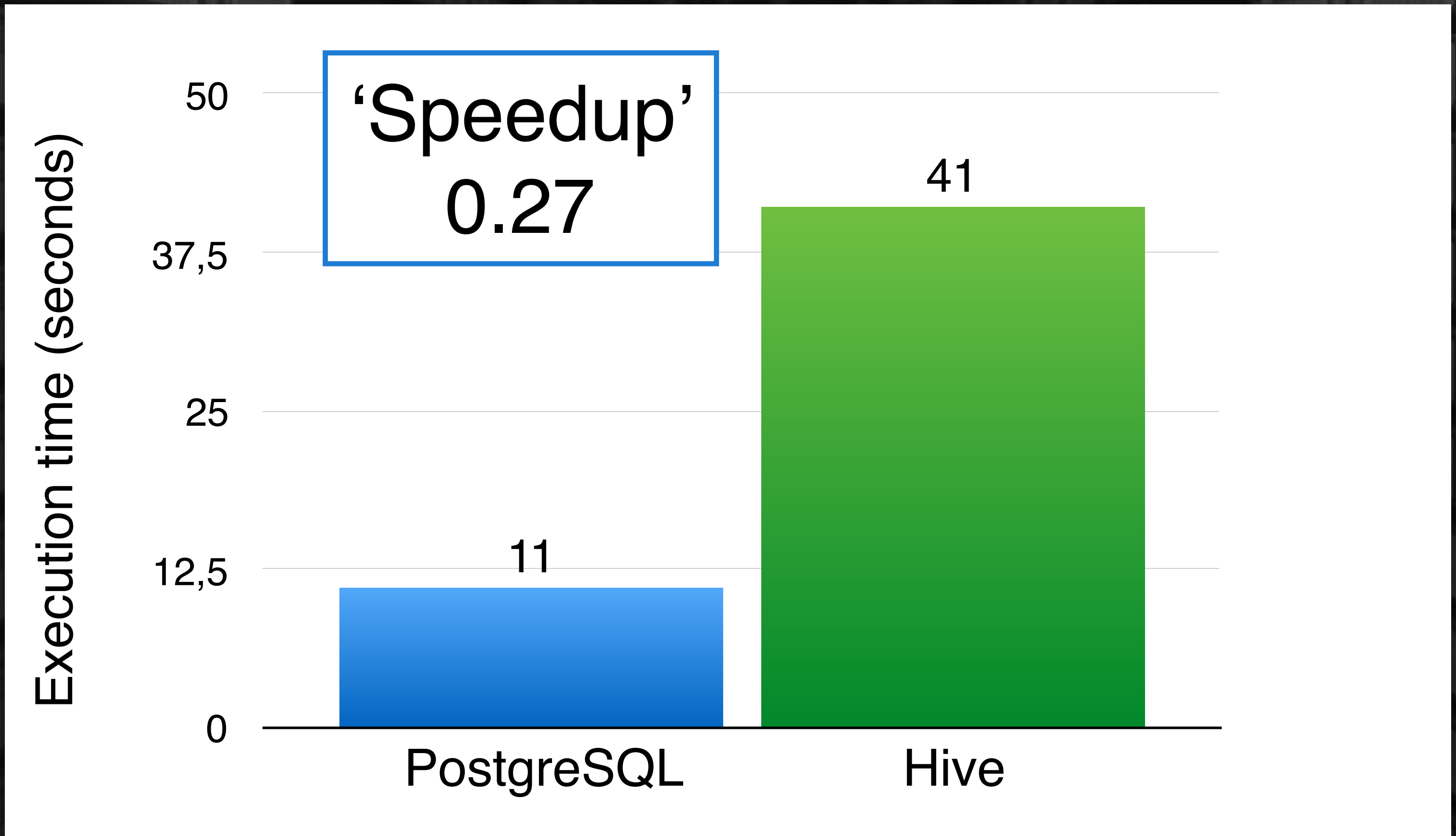
```
SELECT coadd_objects_id, ra, dec, mag_auto_g,
mag_auto_r, mag_auto_i, mag_auto_z,
mean_z_bpz, mode_z_bpz, median_z_bpz,
z_mc_bpz, t_b, spread_model_i,
spreaderr_model_i, modest_class FROM des_y1a1
WHERE                              _i < 22
AND (                              gold = 0)
AND (                              EN -1.
and 3                              i)
BETWE                              i -
mag_a                              (ra < 15
or ra                              out of
~137M
```



'Speedup'
100.0

PostgreSQL vs Hadoop

```
SELECT z, log_m FROM micecatv1 WHERE z < .25
AND z > .23 AND ra < 20 AND dec < 20; (~52K
out of ~205M rows)
```



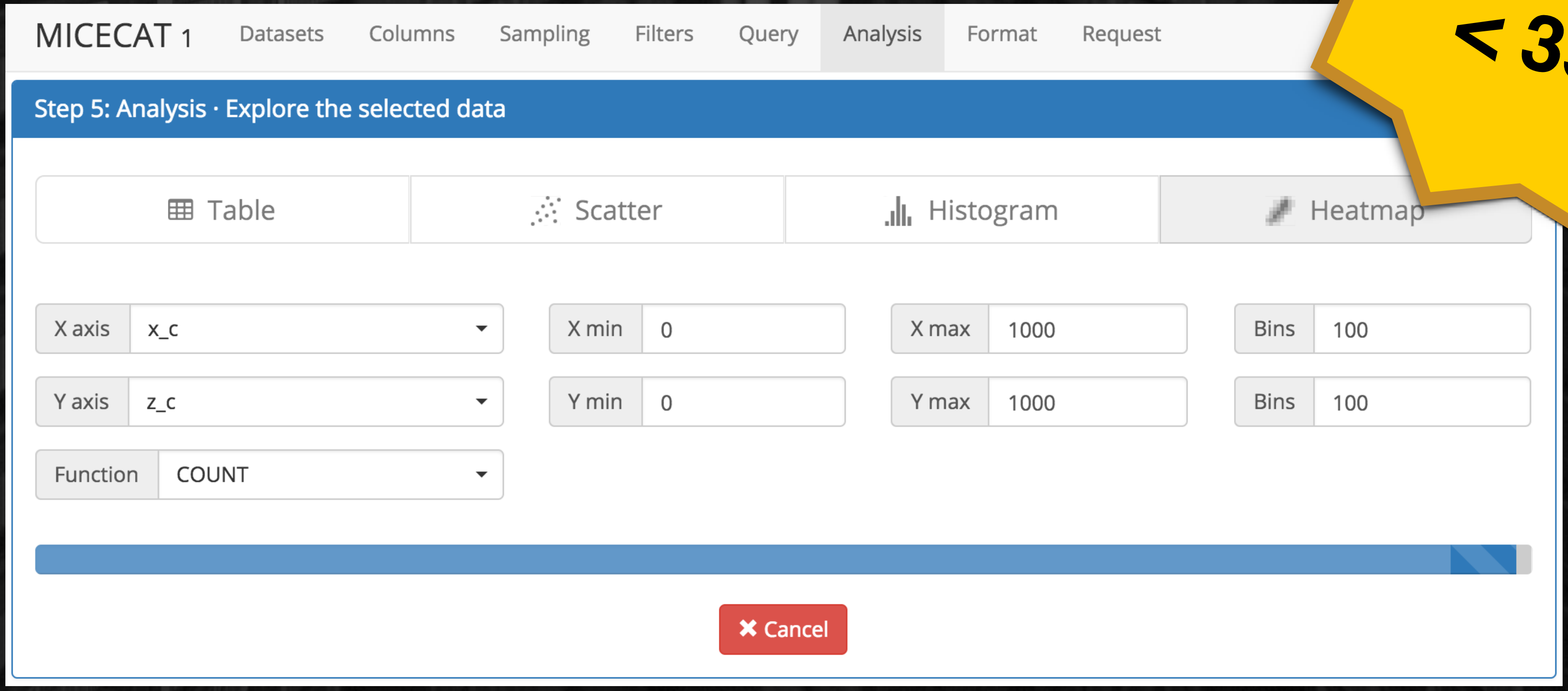
Properly using indices and a very small amount of data requested!

CosmoHub on Hadoop

- CosmoHub is a portal for interactive analysis and distribution of massive cosmology data without any SQL knowledge
- It is built on top of Hadoop and uses the Apache Hive infrastructure
- It is fully developed, hosted and operated at PIC

New features

- Interactive analysis (no time constraint)
- Sampling: select a random subset of the catalog to get faster results when exploring the data
- Heatmap plot
- 2 more file formats to download the selected d FS and ASDF



Step 5: Analysis · Explore the selected data

X axis: x_c — X min: 0 — X max: 1000 — Bins: 100

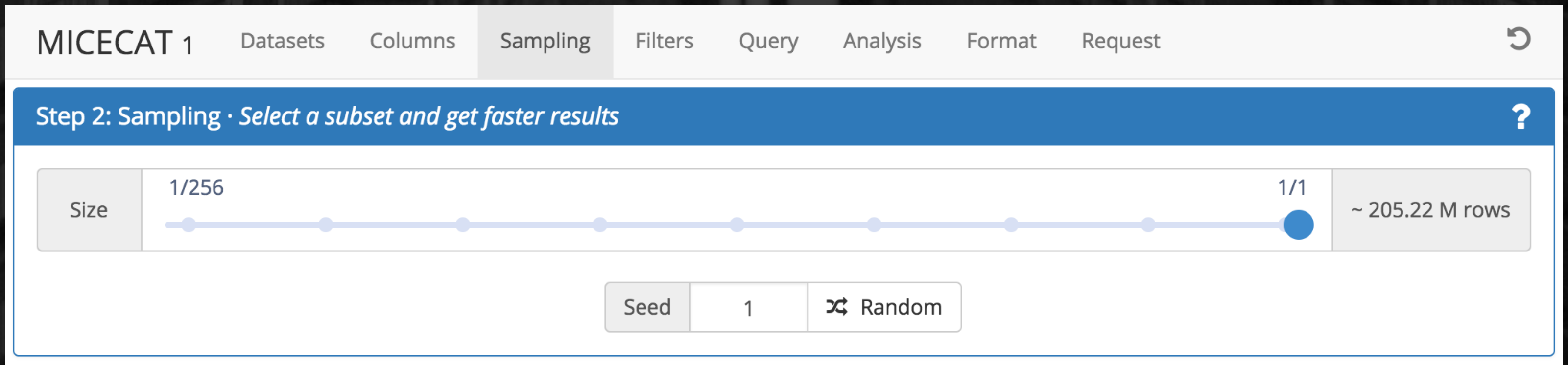
Y axis: z_c — Y min: 0 — Y max: 1000 — Bins: 100

Function: COUNT

500M
< 35''

New features

- Real time analysis (no time constraint)
- Sampling: select a random subset of the catalog to get faster results when exploring the data
- Heatmap plot
- 2 more file formats to download the selected data: FITS and ASDF



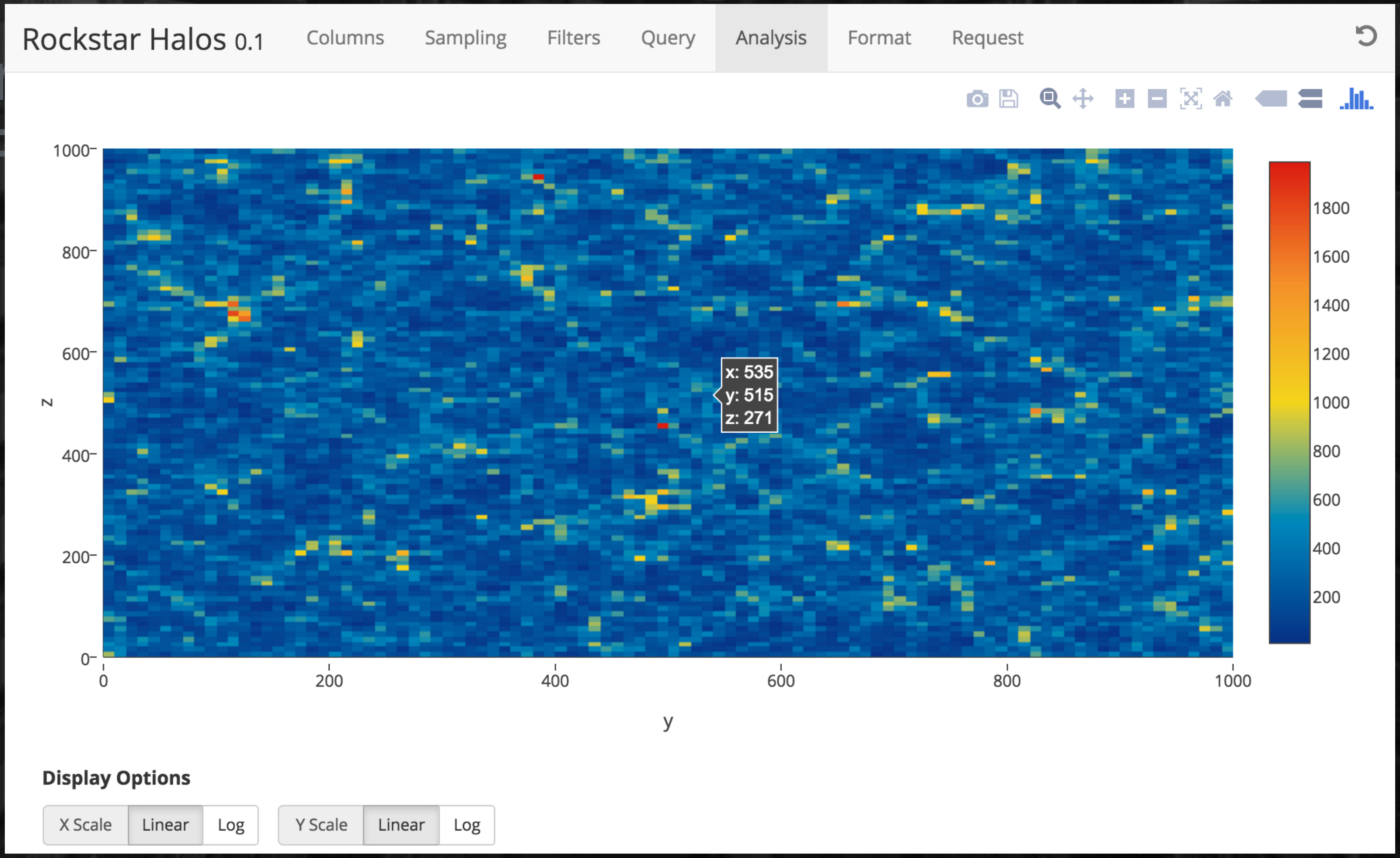
Step 2: Sampling · *Select a subset and get faster results*

Size 1/256 1/1 ~ 205.22 M rows

Seed 1 Random

New features

- Real time analysis (no time constraint)
- Sampling: select a random subset of the catalog to get faster results when exploring the data
- Heatmap plot
- 2 mor FITS and ASDF



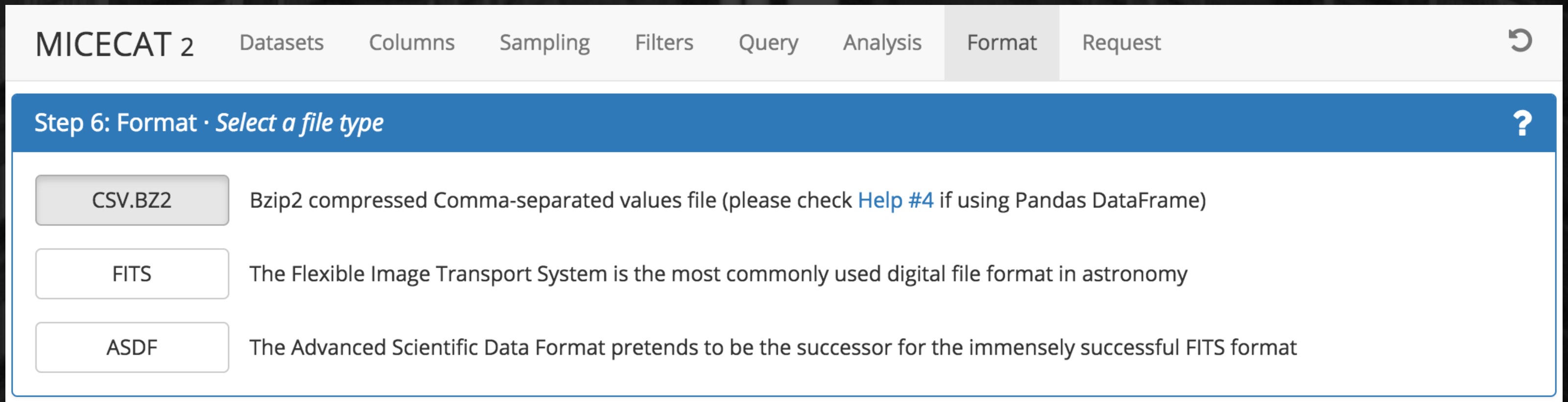
Display Options

X Scale: Linear | Log

Y Scale: Linear | Log

New features

- Real time analysis (no time constraint)
- Sampling: select a random subset of the catalog to get faster results when exploring the data
- Heatmap plot
- 2 more file formats to download the selected data: FITS and ASDF



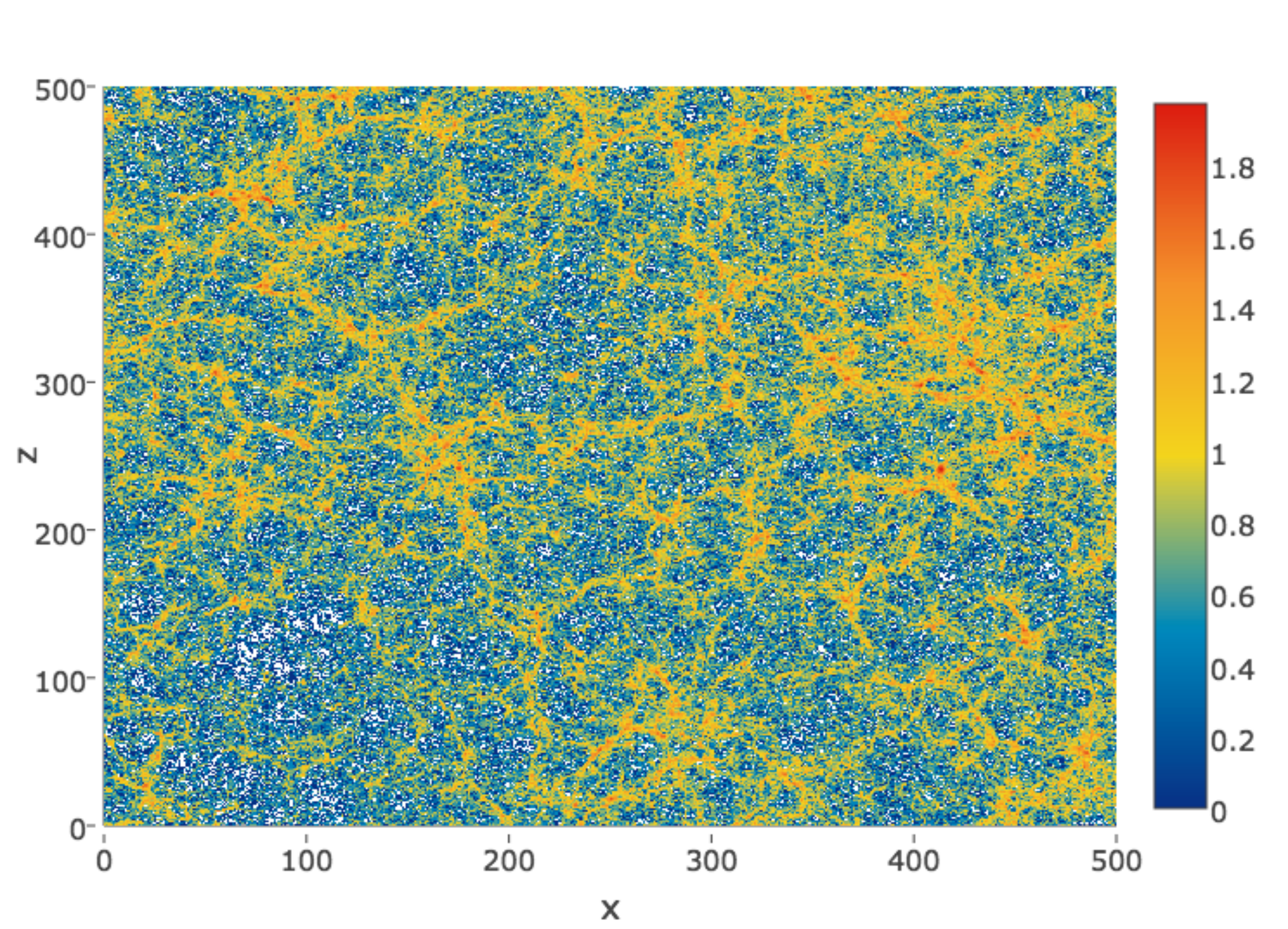
Step 6: Format · *Select a file type*

CSV.BZ2	Bzip2 compressed Comma-separated values file (please check Help #4 if using Pandas DataFrame)
FITS	The Flexible Image Transport System is the most commonly used digital file format in astronomy
ASDF	The Advanced Scientific Data Format pretends to be the successor for the immensely successful FITS format

Demo

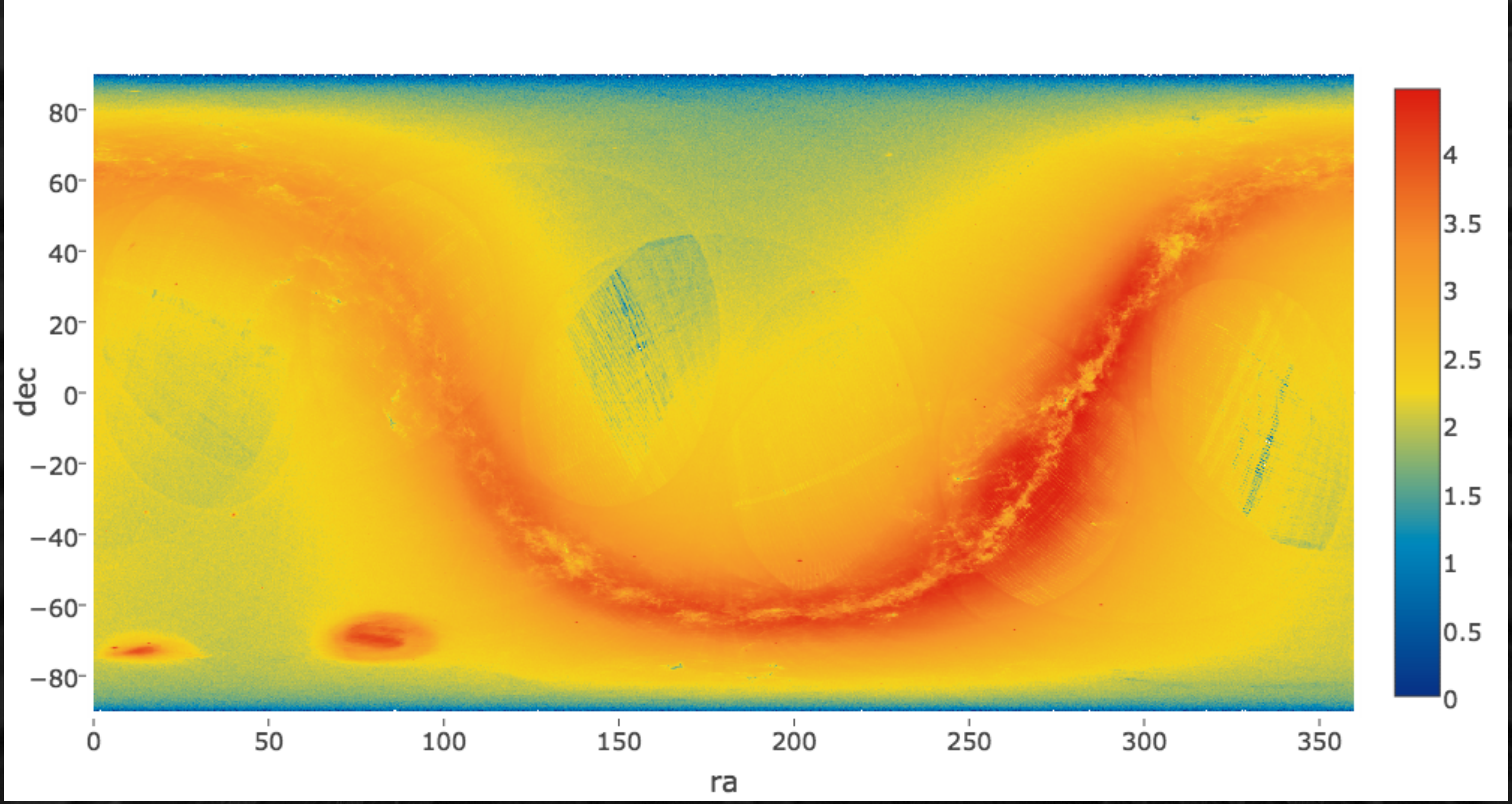
Demo

Euclid Flagship dark matter halo catalog: ~5.5x10^9
500x50x500 Mpc/h section
1000x1000 bins in ~ 4’



Demo

Full Gaia catalog: ~1.1×10^9
ra, dec distribution
1000x1000 bins in ~ 4’



Conclusions & future work

Conclusions

- Great improvement in response time
- New release is more reliable
- Still exploring the vast Hadoop ecosystem

Future work

- New plot types and analysis
- Collaboration with more experiments
  - More data, more catalogs, more users
- Other use cases (other than Cosmology)

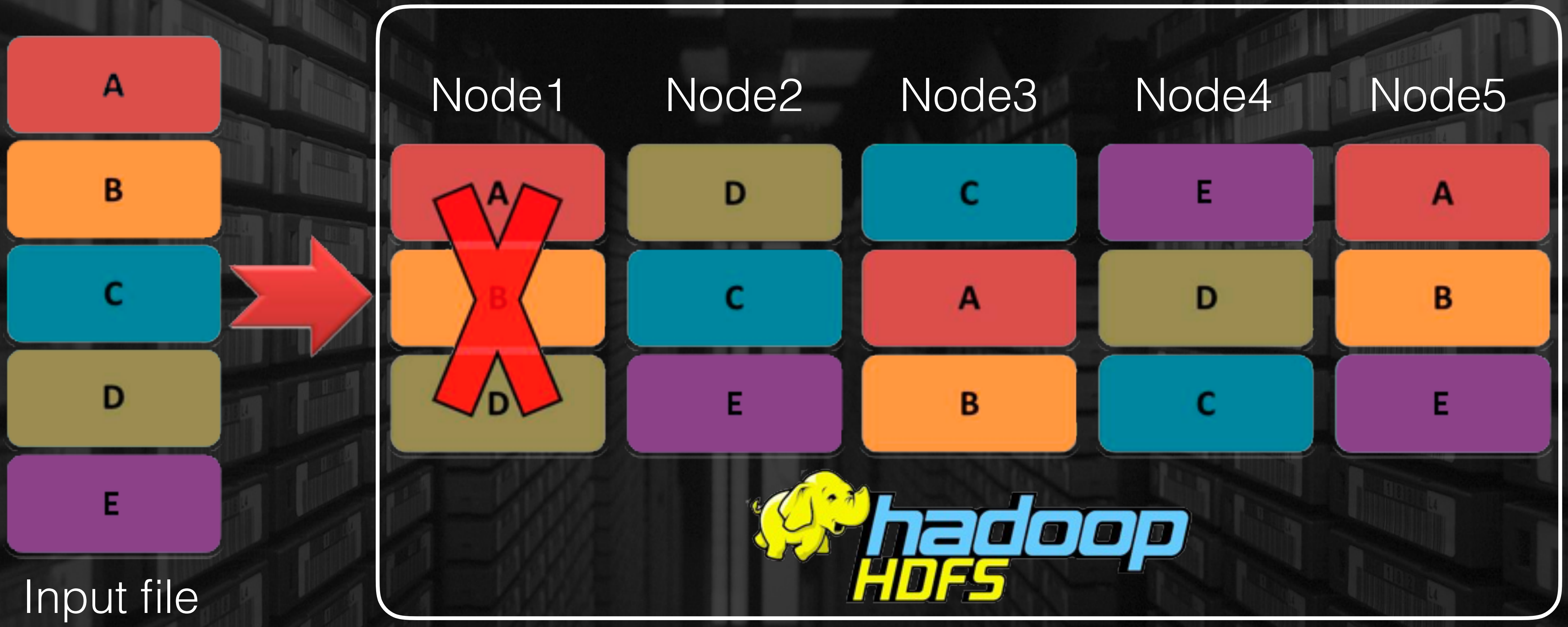
COSMO HUB on Hadoop

https://cosmohub.pic.es

Thanks for your attention!

Backup slides

Hadoop basics



Input file	Node1	Node2	Node3	Node4	Node5
A	A	D	C	E	A
B	B	C	A	D	B
C	D	E	B	C	E
D					
E					

Hive tuning

- We have set the platform so that queries over large tables are really fast:
  - Apache Tez execution engine instead of the venerable Map-reduce engine
  - ORCfile: a new table (column based) storage format
  - Vectorized query technique: batches of 1024 rows at once

Load balancing

- Set up two different queues given the two different profiles:
  - ‘Interactive’: real-time analysis (low latency)
  - ‘Batch’: custom catalogs (high latency)
- Configure queue shares and preemption:
  - batch jobs take idle resources to maximize efficiency (10-90)
  - interactive jobs can take resources from batch queue (90-100)

Backend

- ReST API powered by Flask:
  - flask-restful - ReST framework
  - sqlalchemy - database ORM
  - websockets - bidirectional communications
  - gevent - asynchronous framework
  - pyhive - hive connection library
  - pyhdfs - hdfs bindings

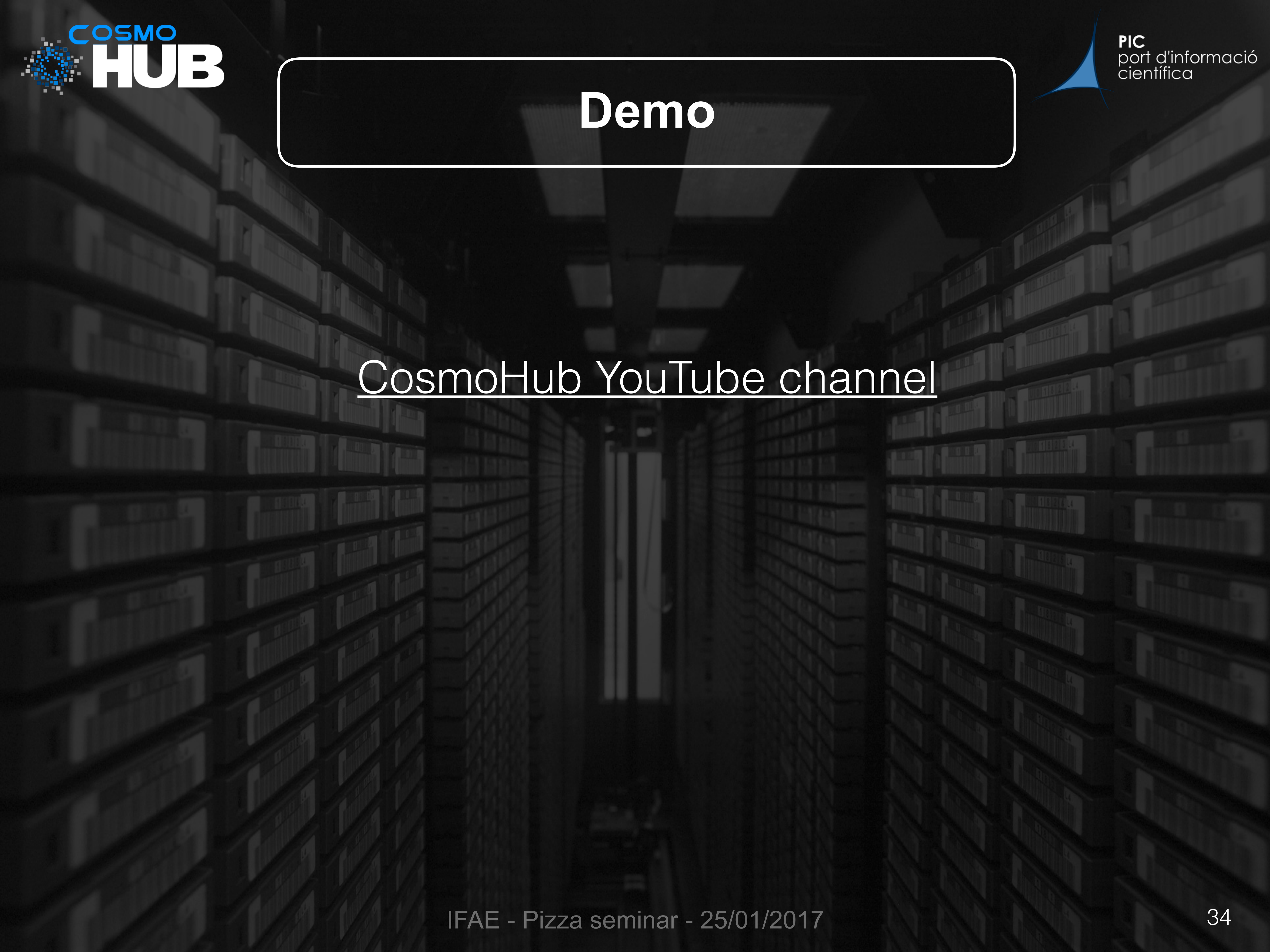
Flask
web development,
one drop at a time

SQLAlchemy

Frontend

- Responsive Web interface powered by:
  - Angular JS - web app oriented HTML framework
  - Bootstrap - responsive frontend framework
  - Plot.ly for plotting
  - Wordpress as backend to edit "static" content

Demo



CosmoHub YouTube channel