PROCEEDINGS OF SCIENCE

CosmoHub and SciPIC: Massive cosmological data analysis, distribution and generation using a Big Data platform

J. Carretero[*a], **P. Tallada**[b], **J. Casals**[b], **M. Caubet**[b], **F. Castander**[c], **L. Blot**[c], **A. Alarcón**[c], **S. Serrano**[c], **P. Fosalba**[c], **C. Acosta-Silva**[a], **N. Tonello**[a], **F. Torradeflot**[a], **M. Eriksen**[a], **C. Neissner**[a] **and M. Delfino**[a]

[a]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology*
Campus UAB, 08193 Bellaterra (Barcelona) Spain

[b]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT)*
Madrid, Spain

[c]*Institute of Space Sciences, IEEC-CSIC, Campus UAB, Carrer de Can Magrans, s/n*
08193 Barcelona, Spain

E-mail: carretero@pic.es, tallada@pic.es, jcasals@pic,es, mcaubet@pic,es, fjc@ieec.uab.es, blot@ice.cat, alarcon@ice.cat, serrano@ieec.uab.es, fosalba@ieec.uab.es, cacosta@pic.es, tonello@pic.es, torradeflot@pic.es, neissner@pic.es *and* delfino@pic.es

Galaxy surveys require support from massive datasets in order to achieve precise estimations of cosmological parameters. The CosmoHub platform (https://cosmohub.pic.es), a web portal to perform interactive analysis of massive cosmological data, and the SciPIC pipeline have been developed at the Port d'Informació Científica (PIC) to provide this support, achieving nearly interactive performance in the processing of multi-terabyte datasets. Cosmology projects currently supported include European Space Agency Euclid space mission, the Dark Energy Survey (DES), the Physics of the Accelerating Universe (PAU) survey and the Marenostrum Institut de Ciències de l'Espai Simulations (MICE). Support for additional projects can be added as needed. CosmoHub enables users to interactively explore and distribute data without any SQL knowledge. It is built on top of Apache Hive, part of the Apache Hadoop ecosystem, which facilitates reading, writing, and managing large datasets. More than 50 billion objects, from public and private data, as well as observed and simulated data, are available. Over 500 registered scientists have produced about 2000 custom catalogs occupying 10TiB in compressed format over the last three years. All those datasets can be interactively explored using an integrated visualization tool. The current implementation allows an interactive analysis of 1.1 billion object datasets to complete in 45 seconds. The SciPIC scientific pipeline has been developed to efficiently generate mock galaxy catalogs using as input a dark matter halo population. It runs on top of the Hadoop platform using Apache Spark, which is an open-source cluster-computing framework. The pipeline is currently being calibrated to populate the full sky Flagship dark matter halo catalog produced by the University of Zürich, which contains about 44 billion dark matter haloes in a box size of 3.78 Gpc/h. The resulting mock galaxy catalog is directly stored in the CosmoHub platform.

PoS(EPS-HEP2017)488

The European Physical Society Conference on High Energy Physics
5-12 July, 2017
Venice

[*] Speaker.

1. Introduction

In the recent years, experimental astronomy has entered into the Big Data data regime mainly due to the incessant construction and development of ground- and space- based sky surveys in the whole electromagnetic spectrum, from gamma rays and X-rays, ultraviolet, optical, and infrared to radio bands, e.g. the future 3.2 gigapixel LSST camera will take images every 30 seconds and the data rate will be about 15 TiB per night [1].

Besides calibrating and reducing the data coming from the telescopes, easily sharing and quickly overviewing and distributing the data is key to know about its veracity and finally to be able to get value from the data. This means that the way scientists access large data volumes has to be revisited. As a result, new interfaces that allow to get to the datasets without downloading them have become necessary. The main problem is not the storage capacity but the time it is going to take the user to get an answer out of the data, so it is very important to choose the proper technology, data formats and software tools to manage the data. Since technologies evolve very rapidly, one can design and develop software and data management tools to be ready for real data and these can quickly become obsolete as new technologies and tools are developed in the meantime.

PostgreSQL solution was not working as we expected for our use cases. As an example, we generated, and ingested into our database the second version of the MICE mock galaxy catalog (MICECAT2) [2], [3] and [4], which contains about 500 million entries and more than 120 fields per entry. The aim of the MICE project is to reproduce, with unprecedented detail, the history of the universe from much before the first stars formed up to nowadays, and to determine how well future astronomical surveys can answer these fundamental open questions. MICECAT2 is composed of several tables that can be efficiently joined. Also, indexes on the most used fields for filtering, such as sky position, redshift or magnitude, were created to ensure fast query response times. However, most queries took a lot of time to finish, up to several hours. This happens because relational databases resort to execute queries through a full table sequential scan when the requested data is larger than 15% of the total volume of the table[1]. In those cases, using indices offers no performance benefit, so they are ignored. Furthermore, other operations such as schema modifications, or removing large subsets of rows are also very time consuming. Finally, due to the non-distributed nature of a PostgreSQL database, there is a limit on the amount of storage space it can handle.

For all these reasons we realized that PostgreSQL was not adequate for our needs and we explored different possibilities to solve the problem. We looked at distributed databases and Big Data platforms, and we settled on one of the most popular Big Data platforms, Hadoop. In this particular work we present two different tools, SciPIC and CosmoHub[2], developed on top of the PIC Big Data platform.

2. Big data platform at PIC

The PIC Big Data platform is based on the Apache Hadoop framework, which enables distributed processing of large datasets. It relies on commodity hardware and it is designed to scale

[1] See https://docs.oracle.com/cd/B12037_01/server.101/b10739/indexes.htm for more info

[2] https://cosmohub.pic.es

up from dozens to thousands of machines, and to automatically detect and recover from failures in a transparent way.

The fundamental technology behind Hadoop is HDFS (Hadoop Distributed File System), which is a highly fault-tolerant distributed file system providing high throughput access for large datasets. HDFS splits data in blocks and distributes them across different data nodes. Like other technologies such like Ceph[3] or GPFS[4], HDFS also makes use of data replication, where a minimum of 3 replicas is recommended for production environments. Thanks to its fault-tolerant and replicated design we are able to extend the HDFS cluster with cheap commodity hardware or recycled machines knowing that, in the event of losing one node, a replica will always be present in a different physical location. Data stored in HDFS can be managed with a HDFS specific client or, alternatively, mounted through NFS[5] and managed with standard tools.

Several applications have been developed on top of HDFS: Apache Hive is a data warehouse that facilitates reading, writing and managing large datasets residing in distributed storage using SQL, and Apache Spark is a distributed processing framework capable of dealing with large volumes of data. All these software components, in addition to administration tools, are provided by the the Hortonworks Data Platform (HDP), which is an open-source Apache Hadoop distribution containing a set of very well tested and integrated Apache Hadoop components. Currently, our Big Data platform at PIC is composed by 30 nodes, with a total of 360 cores, 1080 GiB RAM and 90 TiB HDD.

2.1 SciPIC

SciPIC is a scientific pipeline developed in a collaborative effort between ICE and PIC, to generate mock galaxy catalogs using the Halo Occupation Distribution (HOD) model. This pipeline has been designed to deal with one of the most difficult challenges within the Euclid consortium project: to generate the largest mock galaxy catalog ever using as input the Flagship dark matter N-body simulation produced by the University of Zurich [5]. The dark matter halo catalog used as input contains more than 40 billion dark matter haloes, occupies 5.5TiB in compressed CSV format[6], while covering the full sky up to redshift 2.4.

The pipeline, written in python, is separated in different modules, each one of them in charge of generating a different set of galaxy properties. The resulting galaxy catalog follows some of the most important observed global properties of the galaxy population, such as the luminosity function, the galaxy clustering as a function of the luminosity and color, galaxy shape distribution and lensing properties.

In order to calibrate our algorithm with the Flaship dark matter halo catalog we use SDSS observed data [6]. In particular we use the distribution of galaxies as a function of their luminosity. To achieve this we integrated the 'treecorr' algorithm[7] [7], which estimates 2-point correlation functions, in our Big Data platform. This implementation is useful for many other projects as it is able to deal with very large datasets.

[3] http://ceph.com/publications/

[4] https://dl.acm.org/citation.cfm?id=1083349f

[5] http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.473

[6] https://tools.ietf.org/html/rfc4180

[7] https://github.com/rmjarvis/TreeCorr

The SciPIC pipeline is able to generate one octant of the sky with 2.6 billion galaxies, with 130 properties each, in 5.3 hours. The resulting catalog is accessible through CosmoHub. Figure 1 shows a slice of the 3-dimension light-cone projected from a 40 Mpc/h width in the direction orthogonal to the image plane. The three different panels correspond to different galaxy samples: the top one shows all galaxies in the catalog, the middle and the bottom ones show galaxies which are going to be observed with the VIS instrument[8], and with the NISP channel[9] respectively.

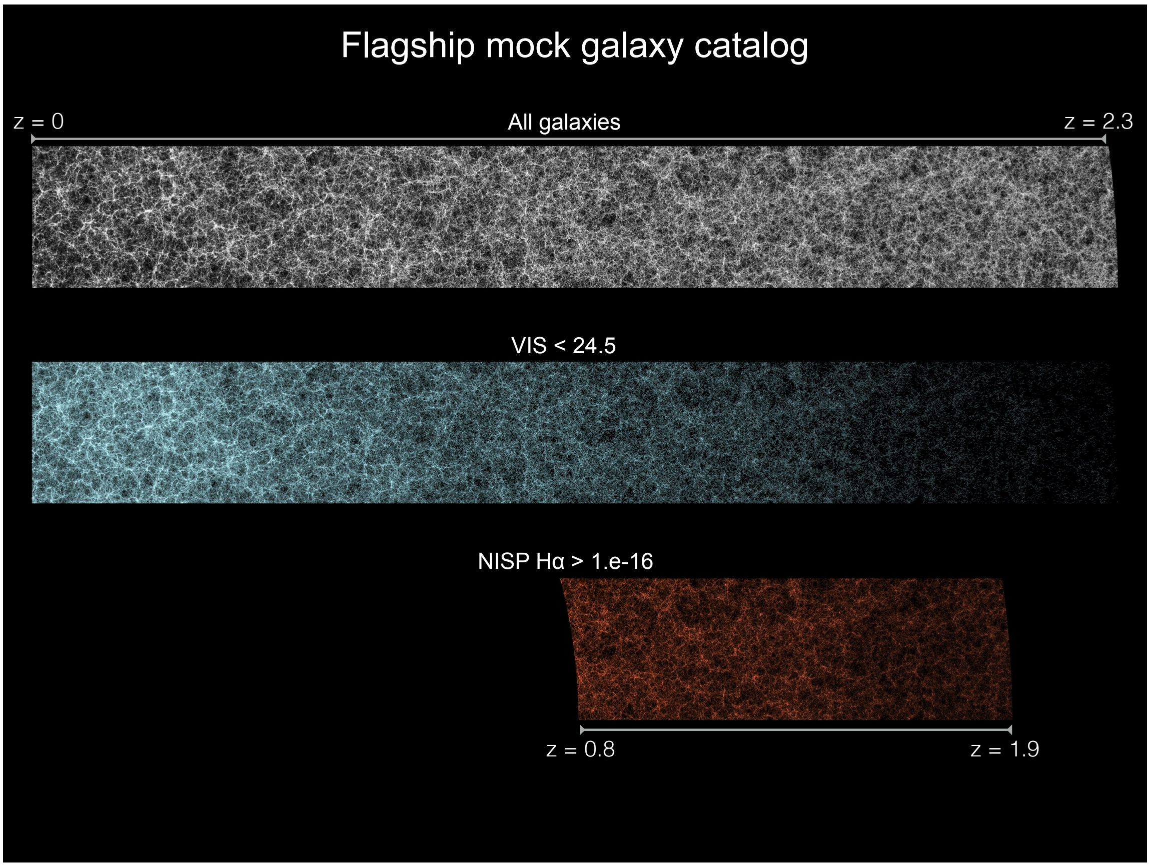

Figure 1: Euclid Flagship mock galaxy catalogue: False colour images showing a small portion (0.3%) of the full light-cone simulation of mock galaxies in the Euclid survey. Light-cone stripes extend 500 Mpc/h (vertical) x 3800 Mpc/h (horizontal axis). The 2-dimension "pencil beam" images result from a slice of the 3-dimension light-cone, projected from a 40 Mpc/h width (in the direction orthogonal to the image plane). From top to bottom, panels display the full sample of galaxies in the mock, and the sub-samples expected from observations in the VIS and NISP-Halpha channels. The galaxy mock has been produced using a Halo Occupation Distribution pipeline developed by the Institut de Ciències de l'Espai (ICE) and Port d'Informació Científica (PIC) in Barcelona, and it is based on the 2 trillion dark-matter particle Flagship run produced by University of Zurich.

2.2 CosmoHub

CosmoHub is a web application to perform interactive exploration and distribution of massive

[8] https://www.euclid-ec.org/?page_id=2485

[9] https://www.euclid-ec.org/?page_id=2490

cosmological data without any SQL knowledge required. It is built on top of Apache Hive, part of the Apache Hadoop ecosystem, which allows users to use a very familiar SQL language and, from the user's perspective, it works similarly as it did before with PostgreSQL. Hive is able to access files stored in HDFS, but also in other data storage systems such like Apache HBase. Figure 2 is a Heatmap plot generated in 45 seconds using "Step 5: Analysis" of the guided process and shows the distribution in the sky of stars in the Milky Way in equatorial coordinates. Data comes from the public first Gaia[10] data release catalog, which contains 1.1 billion objects.

The web portal has been built using a set of widely community supported web development frameworks, such as AngularJS[11] for the HTML and JavaScript[12], Twitter Bootstrap CSS framework[13] for the styling, JavaScript WebSockets[14] for server-client communication, Plot.ly[15] for the charts generated in the Analysis step and Wordpress[16] for some parts of the backend (news, contact form, etc.).

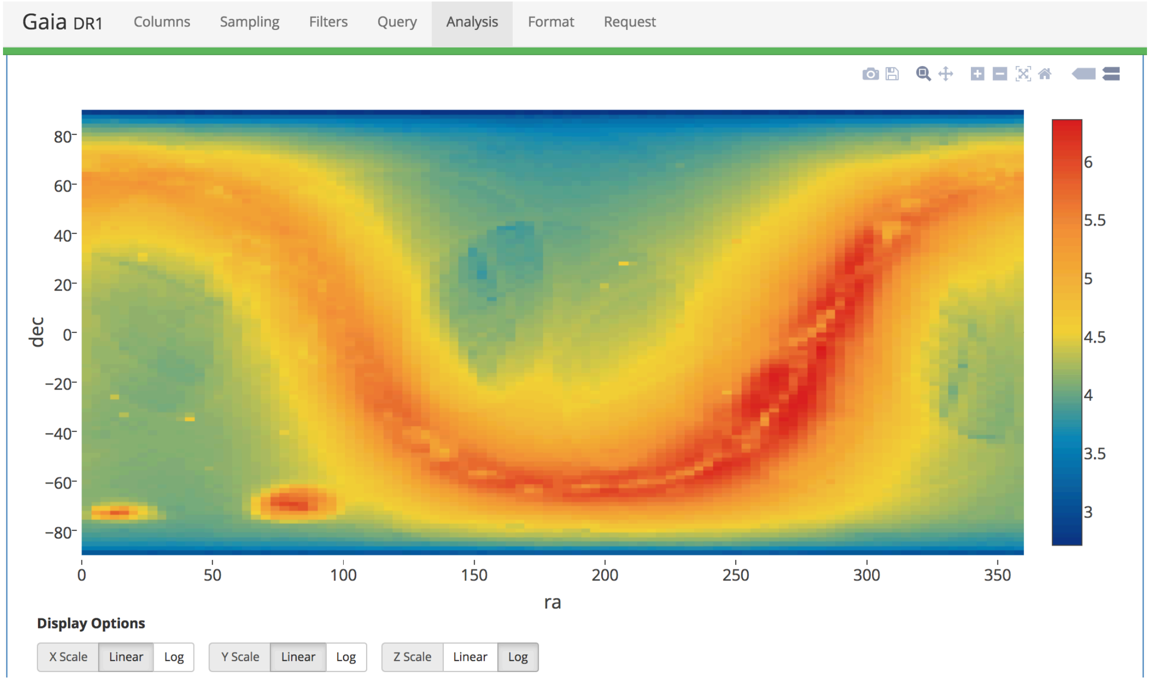

Figure 2: Spatial distribution of stars in our Galaxy in equatorial coordinates. This plot summarizes 1.1 billion stars and has been generated in 35 seconds.

[10] https://www.cosmos.esa.int/web/gaia/dr1-papers

[11] https://angularjs.org/

[12] https://en.wikipedia.org/wiki/JavaScript

[13] http://getbootstrap.com/

[14] https://en.wikipedia.org/wiki/WebSocket

[15] https://plot.ly/

[16] https://en.wikipedia.org/wiki/WordPress

CosmoHub catalogs come from different surveys, which in general produce data of the order of terabytes, which need to be stored, analyzed and distributed. Most of them are galaxy catalogs, but there are also star catalogs such as the first release of the project Gaia[17]). Each entry in a catalog is usually a galaxy, which can be characterized by hundreds of properties (as fields in a table) such as the position in the sky, luminosity, color or morphological properties, etc.

The main feature of CosmoHub is the ability for users to generate their own custom catalogs without any SQL knowledge through a guided process, which can then be analyzed or downloaded in several formats such as CSV.BZ2[18], FITS[19] or ASDF [8]. Users can also download Value-Added-Data, which are *prebuilt* catalogs, sky masks, filter transmission curves, or other files that may be necessary to analyze the data. Catalogs can be interactively explored using an integrated visualization tool which, among others, includes 1-D histograms and 2-D heatmap plots. The visualization tool has been key to explore, validate and distribute the Flagship mock galaxy catalog. Finally, users can use sampling to work on a subset of the total catalog to get even faster results.

3. Conclusions

Relational databases are one of the most powerful tools for dealing with structured data. Well optimized, they can manage hundreds of millions of entries, even some billions, in a single table. They are optimized to handle thousands of queries per second, each one of them involving a small number of records. However, they are not designed to work with queries involving massive subsets of the table. In these cases, a query response may take up to several hours, which precludes any real interactive usage of the results. After migrating the CosmoHub database infrastructure to Apache Hive on top of our PIC Big Data platform, we have experienced a great improvement in the response time: depending on the query, they can run up to 100 times faster.

Migrating to an Apache Hadoop platform enables us to use all the other applications in the ecosystem. For instance, one of these components is the cluster computing framework Apache Spark, which has been key to generate the Euclid Flagship mock galaxy catalog, the largest synthetic galaxy catalog up to date, in just 5 hours.

Finally, we are exploring the possibility of offering the scientists a platform to develop and run complex analysis over our Big Data platform. Projects such as Jupyter[20] or Apache Zeppelin[21] provide such a development and execution environment on top of Apache Spark, and may open a future full of opportunities never foreseen.

References

[1] LSST Dark Energy Science Collaboration LSSTpaper *Large Synoptic Survey Telescope: Dark Energy Science Collaboration*, *ArXiv e-prints*, [arXiv:1211.0310]

[17] https://www.cosmos.esa.int/web/gaia/release

[18] https://en.wikipedia.org/wiki/Comma-separated_values, https://en.wikipedia.org/wiki/Bzip2

[19] https://fits.gsfc.nasa.gov/

[20] http://jupyter.org/

[21] https://zeppelin.apache.org/

[2] J. Carretero et al. *An algorithm to build mock galaxy catalogues using MICE simulations*, *MNRAS*, year = 2015, month = feb, volume = 447, pages = 646-670, doi = 10.1093/mnras/stu2402, [arXiv:1411.3286]

[3] Fosalba et al. *The MICE grand challenge lightcone simulation - I. Dark matter clustering*, *MNRAS*, year = 2015, month = apr, volume = 448, pages = 2987-3000, doi = 10.1093/mnras/stv138, [arXiv:1312.1707]

[4] Crocce et al. *The MICE Grand Challenge lightcone simulation - II. Halo and galaxy catalogues*, *MNRAS*, year = 2015, month = oct, volume = 453, pages = 1513-1530, doi = 10.1093/mnras/stv1708, [arXiv:1312.2013]

[5] Potter et al. *PKDGRAV3: Beyond Trillion Particle Cosmological Simulations for the Next Era of Galaxy Surveys*, *Computational Astrophysics and Cosmology*, year = 2017, month = may, doi = 10.1186/s40668-017-0021-1

[6] Blanton et al. *The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1*, *ApJ*, year = 2003, month = aug, volume = 592, pages = 819-838, doi = 10.1086/375776, [arXiv:astro-ph/0210215]

[7] Jarvis et al. *The skewness of the aperture mass statistic*, *MNRAS*, year = 2004, month = jul, volume = 352, pages = 338-352, doi = 10.1111/j.1365-2966.2004.07926.x, [arXiv:astro-ph/0307393]

[8] Greenfield et al. *ASDF: A new data format for astronomy*, *Astronomy and Computing*, year = 2015, month = sep, volume = 12, pages = 240-251, doi = 10.1016/j.ascom.2015.06.004