



CERN  
openlab

# Data Analytics

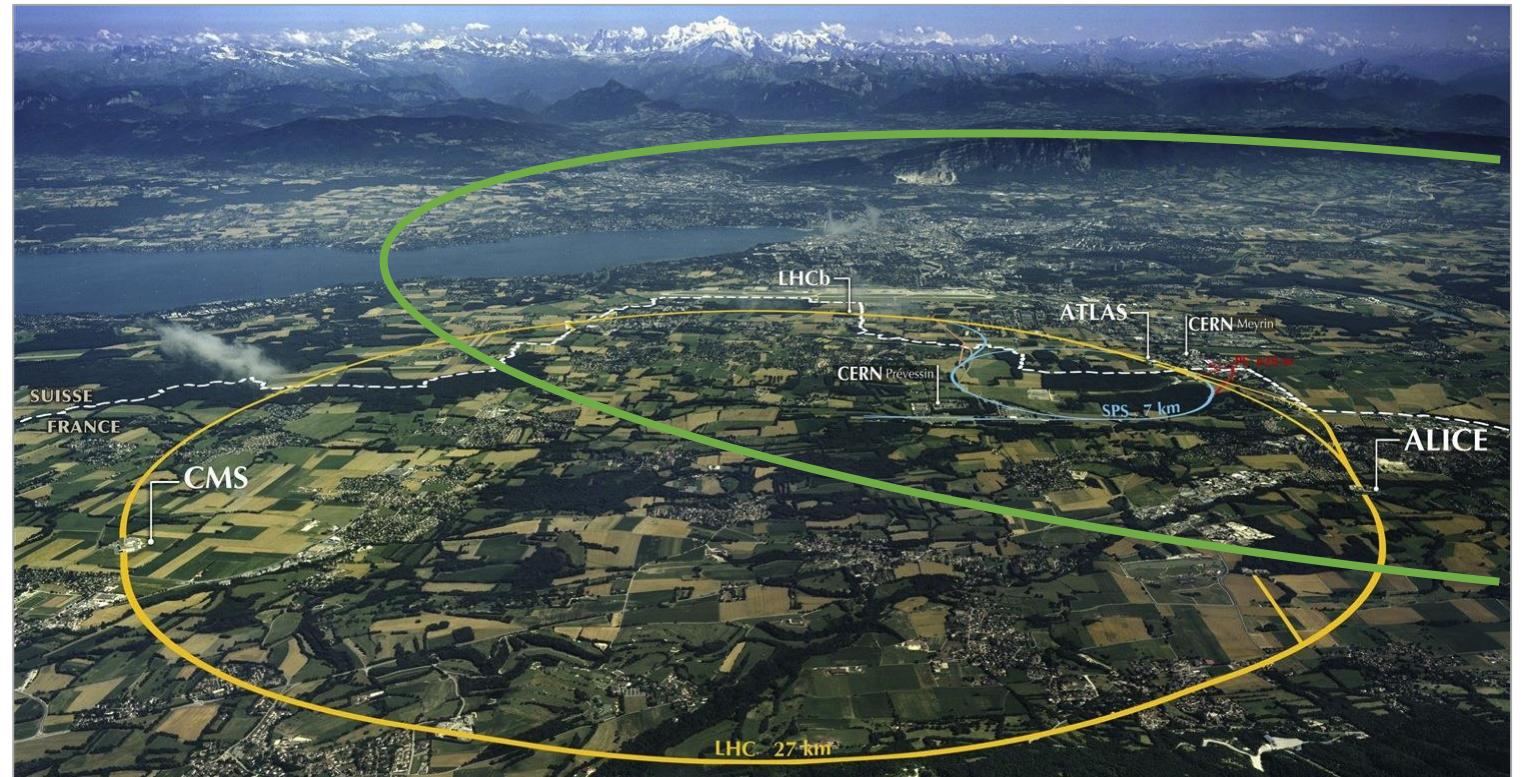
*CERN openlab Open Day*

Manuel Martin Marquez

# Self-Service Data Analytics

## *Future Circular Collider Study*

- Collaboration with more than 70 institutes from all over the world
- Need for a central data repository, which integrates several data sources
  - Control Configuration
  - Fault tracking data
  - Control lot Systems
  - Operation Logbook
- Leverage analytics capabilities for highly heterogenous objectives
- Different expertise and backgrounds involved

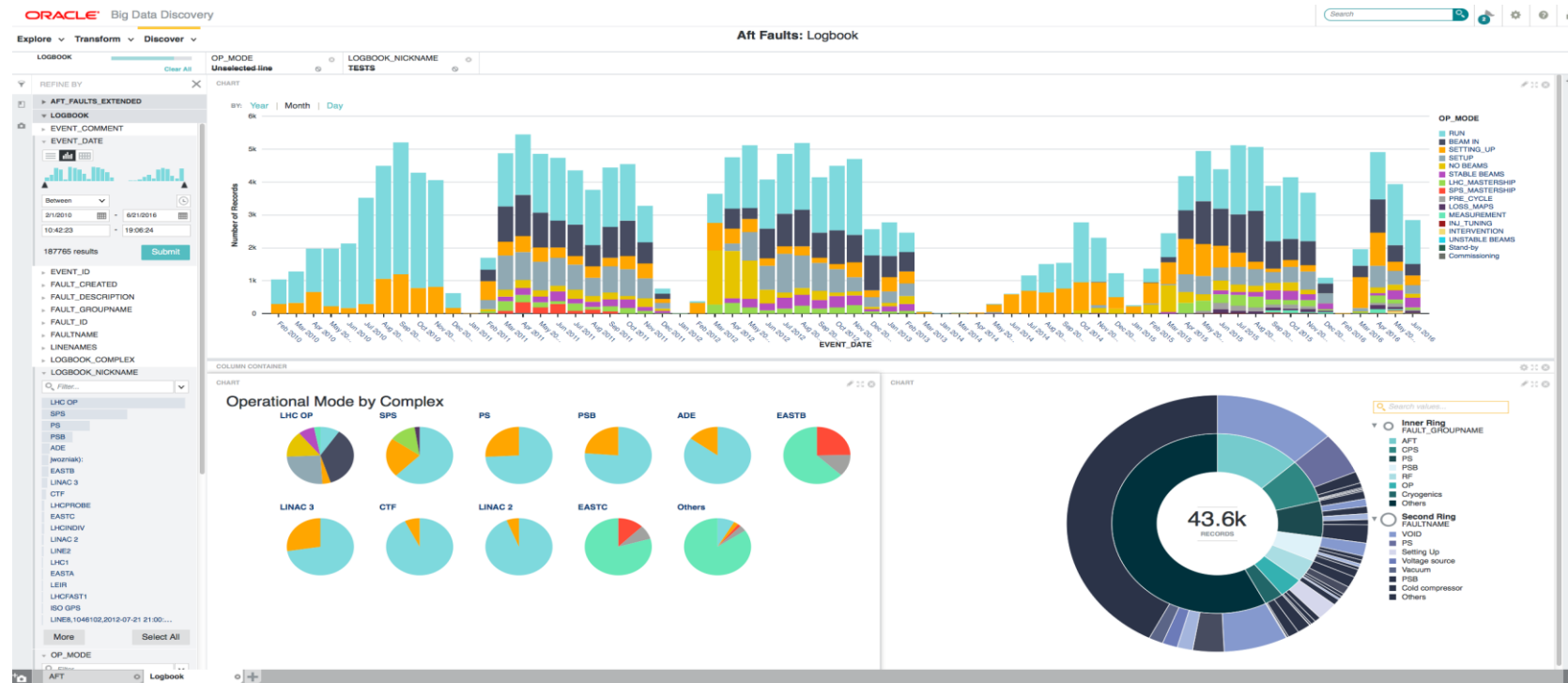


# Self-Service Data Analytics

## Future Circular Collider – Oracle Big Data Discovery

- Agile data governance
- User driven ETL
- Hide technology complexity

- Self-Service discovery tools
- Collaboration and findings sharing
- Integrate seamlessly with advance analytics



# Self-Service Data Analytics

## Graph Data Bases – Oracle Parallel Graph Analytics

- Assess the Data Analytics advantages of Graph Databases
- Find patterns and indirect relations not easy to obtain with traditional database systems
- Evaluate the potential applications of the technology within CERN

- Proof-of-Concepts:
  - Zenodo
  - Accelerator Faults analysis

The screenshot shows a Zenodo record for the paper "Introducing Parsl: A Python Parallel Scripting Library". The record includes the following information:

- Publication date:** August 30, 2017
- DOI:** 10.5281/zenodo.891533
- Keywords:** Parallel scripting, Parsl, Python
- Communities:** Zenodo
- License (for files):** Creative Commons Attribution 4.0
- Versions:** Version 2 (10.5281/zenodo.891533) and Version 1 (10.5281/zenodo.805492)
- Cite all versions:** You can cite all versions by using the DOI 10.5281/zenodo.853491. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)
- Share:** Buttons for Facebook, Twitter, LinkedIn, and a plus sign for more options.
- Cite as:** Babuji, Yadu, Brizius, Alison, Chard, Kyle, Foster, Ian, Katz, Daniel S., Wilde, Michael, & Wozniak, Justin. (2017, August 30). Introducing Parsl: A Python Parallel Scripting Library. Zenodo. <http://doi.org/10.5281/zenodo.891533>

The main content area shows a preview of the paper's abstract and introduction. The abstract states: "Abstract—Researchers frequently rely on large-scale and domain-specific workflows to conduct their science. These workflows may integrate a variety of independent software functions and external applications. However, developing and executing such workflows can be difficult, requiring complex orchestration and management of applications and data as well as customization for specific execution environments. Parsl (Parallel Scripting Library), a Python library for programming and executing data-oriented workflows in parallel, addresses these problems. Developers simply annotate a Python script with Parsl directives; Parsl manages the execution of the script on clusters, clouds, grids, and other resources. Parsl orchestrates required data movement and manages the execution of Python functions and external applications in parallel. In this abstract we describe Parsl's architecture and highlight two domains in which it has been used."

# Self-Service Data Analytics

## Some Future Plans – Cloud Analytics

- Oracle Data Visualization and Discovery **Cloud** solutions
- Streaming Data – Kafka, Oracle Stream Explorer and Golden Gate
- Push analytics computation to the cloud
- Evaluate **services integration**

