

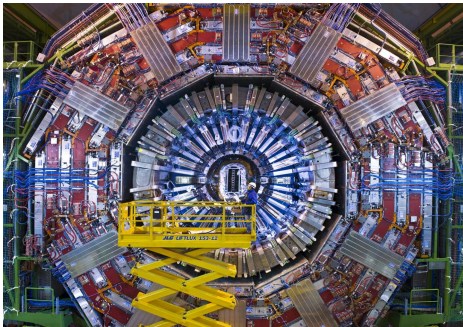
On-Demand Distributed Workflows for CMS Physics Analysis

Diyaselis Delgado López
University of Puerto Rico Mayagüez



UPR
Recinto Universitario de Mayagüez

M UNIVERSITY OF MICHIGAN



- The Compact Muon Solenoid (CMS) is a general-purpose detector at the LHC.
- CMS is a strong contributor to the CERN Open Data Portal
- CMS releases and distributes through CODP reconstructed data, simulations and software

opendata
CERN

is the access point to data produced in research done at CERN

reana

is a reusable and reproducible research data analysis platform.

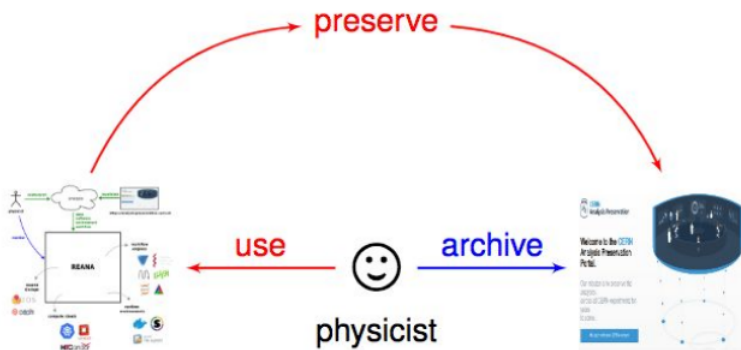
- Structures input data, analysis code, containerized environments and computational workflows



CERN
ANALYSIS PRESERVATION

is a central platform for the four LHC collaborations

- Developed to address the need for the long-term preservation of the data analysis process



REANA

CERN Analysis Preservation

reuse

Higgs-to-four-lepton analysis example using 2011-2012 data

Simplified reimplementaion of the Higgs discovery: Demo

- Uses CMS Open Data inputs and available CMS Software environment
- ReANA, captures the analysis workflow and runs the commands to obtain the output

```
#!/usr/bin/env bash
#SBATCH --job-name=higgs
#SBATCH --output=higgs.out
#SBATCH --error=higgs.err
#SBATCH --time=01:00:00
#SBATCH --mem=1000000000

# Environment
export PYTHONPATH=$(python -c 'import sys; print(sys.path[0])')

# Imports
import os
import sys
import subprocess
import glob
import re
import json
import math
import logging
import argparse
import itertools
import random
import numpy as np
import pandas as pd
import ROOT

# Parameters
parser = argparse.ArgumentParser()
parser.add_argument('--input', type=str, required=True, help='Input directory')
parser.add_argument('--output', type=str, required=True, help='Output directory')
parser.add_argument('--mc', type=bool, default=False, help='Monte Carlo simulation')
parser.add_argument('--year', type=int, default=2011, help='Year (2011 or 2012)')
parser.add_argument('--channel', type=str, default='ee', help='Channel (ee or muu)')
parser.add_argument('--mass', type=float, default=125, help='Higgs mass [GeV]')
parser.add_argument('--width', type=float, default=4, help='Higgs width [GeV]')
parser.add_argument('--scale', type=float, default=1, help='Signal scale factor')
parser.add_argument('--seed', type=int, default=42, help='Random seed')
args = parser.parse_args()

# Paths
input_dir = args.input
output_dir = args.output
mc = args.mc
year = args.year
channel = args.channel
mass = args.mass
width = args.width
scale = args.scale
seed = args.seed

# Setup
os.makedirs(output_dir, exist_ok=True)
random.seed(seed)

# Data
def get_data():
    """Get data from input directory"""
    files = glob.glob(os.path.join(input_dir, '*.*'))
    data = []
    for file in files:
        # Read file content
        with open(file, 'r') as f:
            content = f.read()
        # Parse content
        data.append(parse_content(content))
    return data

def parse_content(content):
    """Parse content of a file"""
    # Example: parse a JSON object
    try:
        obj = json.loads(content)
    except:
        # Example: parse a text line
        obj = {}
    return obj

# Analysis
def analyze_data(data):
    """Analyze the data"""
    # Example: calculate signal yield
    yield = 0
    for event in data:
        # Example: check if event is signal
        if event['channel'] == channel:
            yield += 1
    return yield

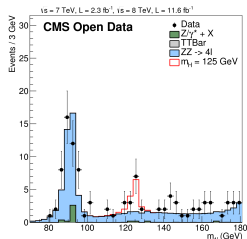
# Main
if __name__ == '__main__':
    # Get data
    data = get_data()

    # Analyze data
    yield = analyze_data(data)

    # Save results
    with open(os.path.join(output_dir, 'results.json'), 'w') as f:
        json.dump({'yield': yield}, f)
```

Progress and Next Steps

- Test CWL workflows locally
- Run analysis inside Docker container locally
- Run analysis within the ReANA platform
(Waiting for Upgrade)



- CERN Open Data Portal Validation Examples: Reprocessing AOD from 2010-2012 RAW samples
- CMS Data Analysis School Example: Top quark mass measurement from b-jet energy spectrum



Paris · Annecy · Chamonix