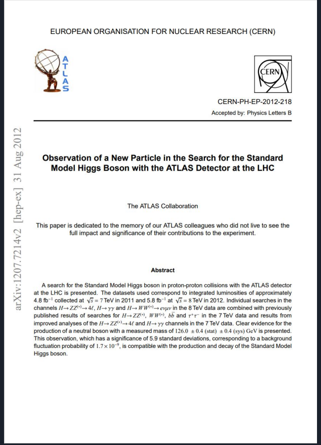
Using Python, coffea, and ServiceX to Rediscover the Higgs. Twice.

Baidyanath Kundu (Manipal Institute of Technology)

Gordon Watts (University of Washington)

What Higgs, Exactly?



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218

Accepted by: Physics Letters B

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012

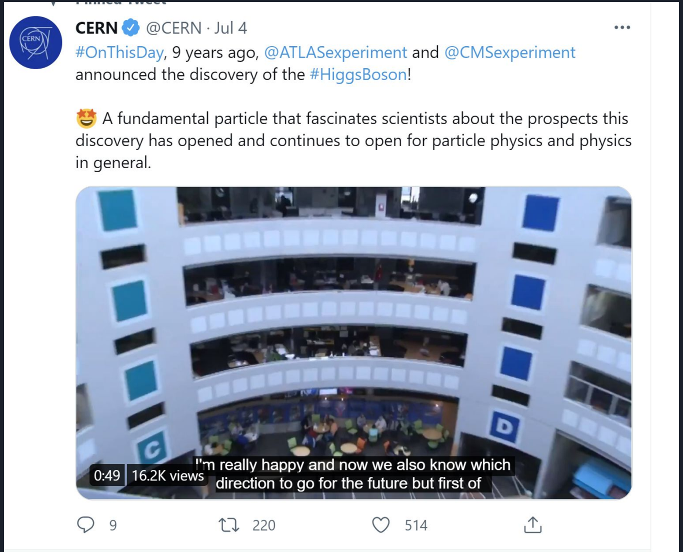
Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

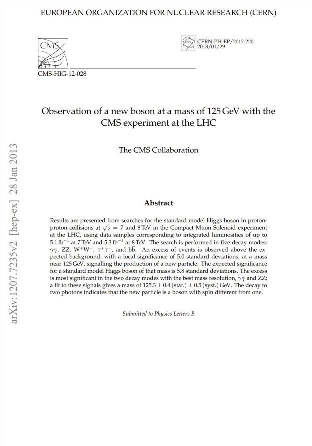


CERN @CERN · Jul 4

#OnThisDay, 9 years ago, @ATLASexperiment and @CMSexperiment announced the discovery of the #HiggsBoson!

🤩 A fundamental particle that fascinates scientists about the prospects this discovery has opened and continues to open for particle physics and physics in general.

0:49
16.2K views
I'm really happy and now we also know which direction to go for the future but first of
9
220
514



EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP/2012-220
2013/01/29

CMS-HIG-12-028

arXiv:1207.7235v2 [hep-ex] 28 Jan 2013

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

The CMS Collaboration

Abstract

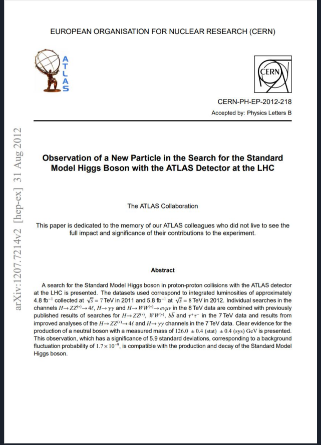
Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the Compact Muon Solenoid experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, $\mathrm{W^+W^-}$, $\tau^+\tau^-$, and $\mathrm{b\bar{b}}$. An excess of events is observed above the expected background, with a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ; a fit to these signals gives a mass of $125.3 \pm 0.4\,(\mathrm{stat.}) \pm 0.5\,(\mathrm{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

Submitted to Physics Letters B

Both experiments have released their Run 1 Higgs Discovery Data on CERN Open Data

Including Source Code To Reproduce Analyses!

What Higgs, Exactly?



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218

Accepted by: Physics Letters B

arXiv:1207.7214v2 [hep-ex] 31 Aug 2012

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

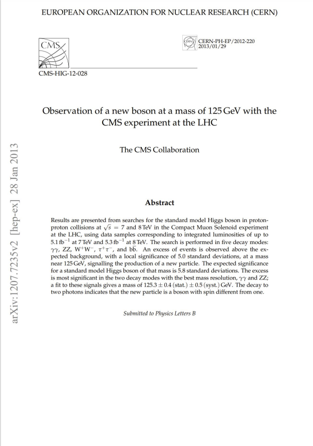
The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Can we reproduce their results using the ServiceX, func_adl, coffea tool chain?



EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP/2012-220
2013/01/29

CMS-HIG-12-028

arXiv:1207.7235v2 [hep-ex] 28 Jan 2013

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

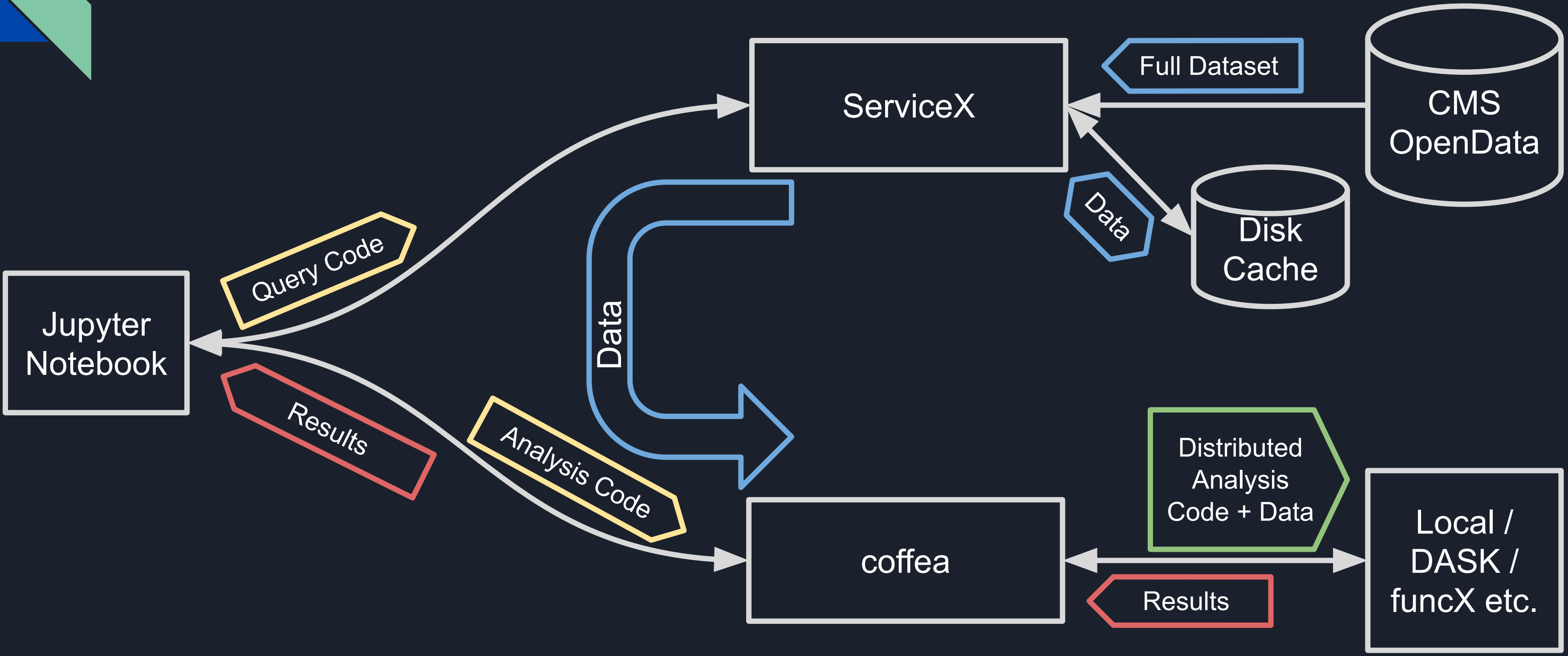
The CMS Collaboration

Abstract

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the Compact Muon Solenoid experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, W$^+$W$^-$, $\tau^+\tau^-$, and b$\bar{\mathrm{b}}$. An excess of events is observed above the expected background, with a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ; a fit to these signals gives a mass of $125.3 \pm 0.4\,(\mathrm{stat.}) \pm 0.5\,(\mathrm{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

Submitted to Physics Letters B

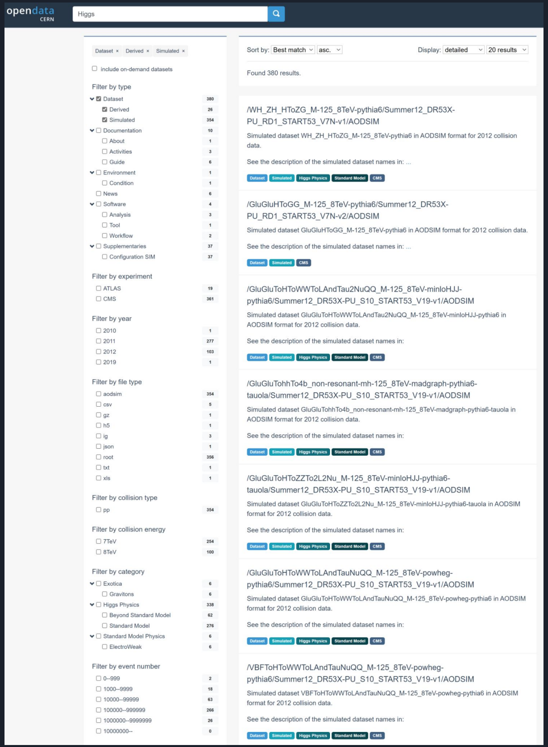
Analysis In Python



CERN opendata

opendata
CERN

Over 2 PB of data, 1000’s of datasets



Higgs

Dataset × Derived × Simulated ×

☐ include on-demand datasets

Filter by type

☑ Dataset	380
☑ Derived	26
☑ Simulated	354
☐ Documentation	10
☐ About	1
☐ Activities	3
☐ Guide	6
☐ Environment	1
☐ Condition	1
☐ News	6
☐ Software	4
☐ Analysis	3
☐ Tool	1
☐ Workflow	2
☐ Supplementaries	37
☐ Configuration SIM	37

Filter by experiment

☐ ATLAS	19
☐ CMS	361

Filter by year

☐ 2010	1
☐ 2011	277
☐ 2012	103
☐ 2019	1

Filter by file type

☐ aodsim	354
☐ csv	5
☐ gz	1
☐ h5	1
☐ ig	3
☐ json	1
☐ root	356
☐ txt	1
☐ xls	1

Filter by collision type

☐ pp	354

Filter by collision energy

☐ 7TeV	254
☐ 8TeV	100

Filter by category

☐ Exotica	6
☐ Gravitons	6
☐ Higgs Physics	338
☐ Beyond Standard Model	62
☐ Standard Model	276
☐ Standard Model Physics	6
☐ ElectroWeak	6

Filter by event number

☐ 0--999	2
☐ 1000--9999	18
☐ 10000--99999	63
☐ 100000--999999	266
☐ 1000000--9999999	26
☐ 10000000--	0

Sort by: Best match asc. Display: detailed 20 results

Found 380 results.

/WH_ZH_HToZG_M-125_8TeV-pythia6/Summer12_DR53X-PU_RD1_START53_V7N-v1/AODSIM

Simulated dataset WH_ZH_HToZG_M-125_8TeV-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in: ...

Dataset Simulated Higgs Physics Standard Model CMS

/GluGluHToGG_M-125_8TeV-pythia6/Summer12_DR53X-PU_RD1_START53_V7N-v2/AODSIM

Simulated dataset GluGluHToGG_M-125_8TeV-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in: ...

Dataset Simulated CMS

/GluGluToHToWWToLAndTau2NuQQ_M-125_8TeV-minloHJJ-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Simulated dataset GluGluToHToWWToLAndTau2NuQQ_M-125_8TeV-minloHJJ-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in:

Dataset Simulated Higgs Physics Standard Model CMS

/GluGluTohhTo4b_non-resonant-mh-125_8TeV-madgraph-pythia6-tauola/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Simulated dataset GluGluTohhTo4b_non-resonant-mh-125_8TeV-madgraph-pythia6-tauola in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in:

Dataset Simulated Higgs Physics Standard Model CMS

/GluGluToHToZZTo2L2Nu_M-125_8TeV-minloHJJ-pythia6-tauola/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Simulated dataset GluGluToHToZZTo2L2Nu_M-125_8TeV-minloHJJ-pythia6-tauola in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in:

Dataset Simulated Higgs Physics Standard Model CMS

/GluGluToHToWWToLAndTauNuQQ_M-125_8TeV-powheg-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Simulated dataset GluGluToHToWWToLAndTauNuQQ_M-125_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in:

Dataset Simulated Higgs Physics Standard Model CMS

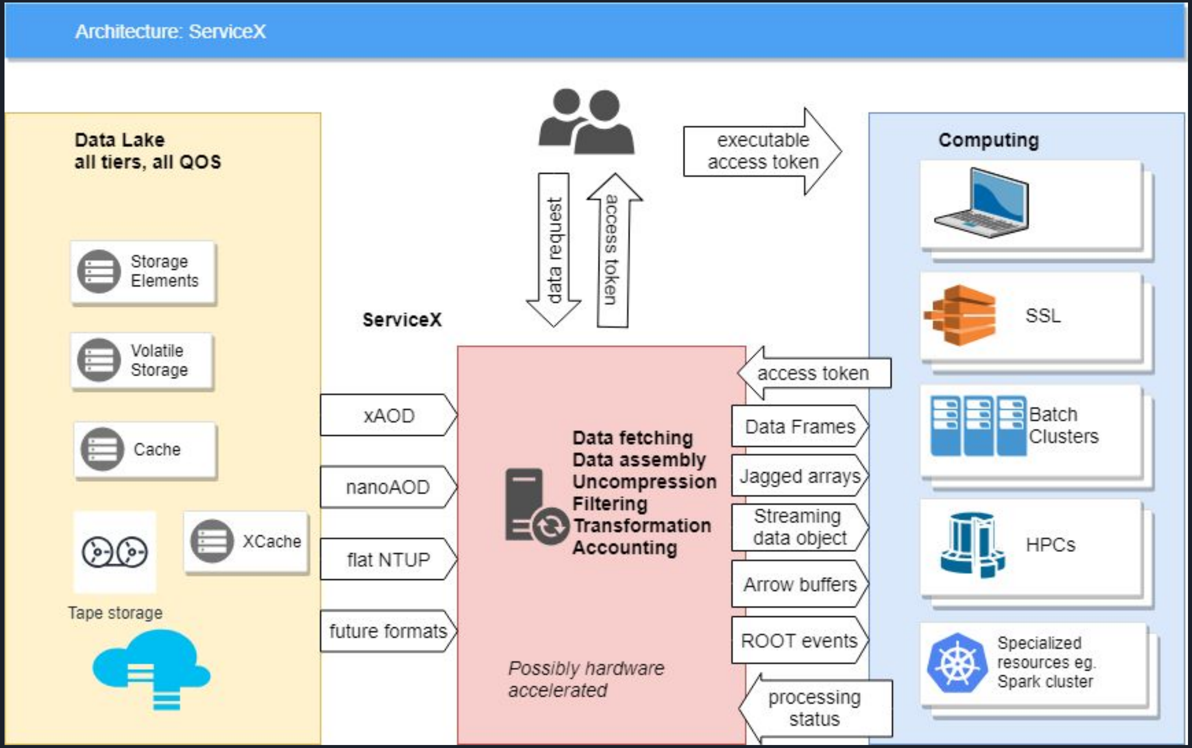
/VBFToHToWWToLAndTauNuQQ_M-125_8TeV-powheg-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM

Simulated dataset VBFToHToWWToLAndTauNuQQ_M-125_8TeV-powheg-pythia6 in AODSIM format for 2012 collision data.

See the description of the simulated dataset names in:

Dataset Simulated Higgs Physics Standard Model CMS

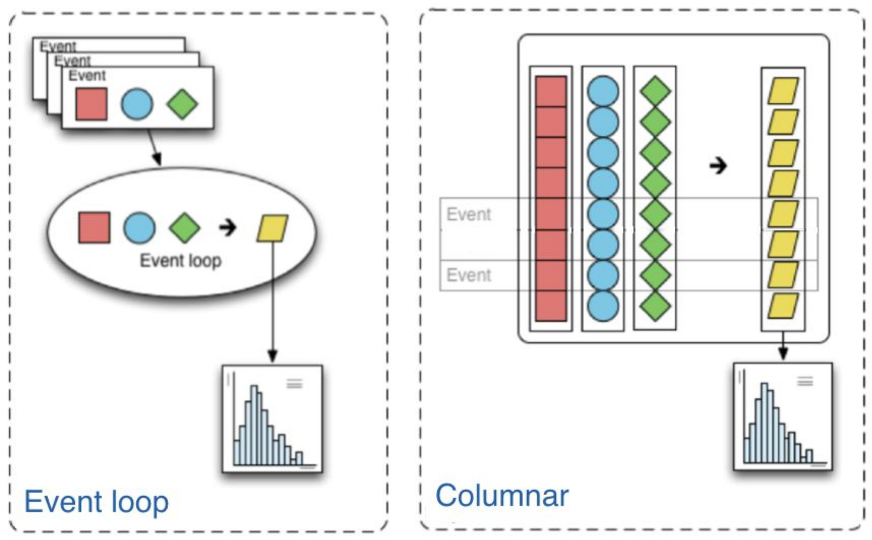
ServiceX: Overview



coffea

A light-weight processing framework for processing event columnar data.

- Processing files
- Distributed processing in a farm (DASK, slurm, Spark, etc.)
- Flexible data model that makes column-wise data look like row-wise.



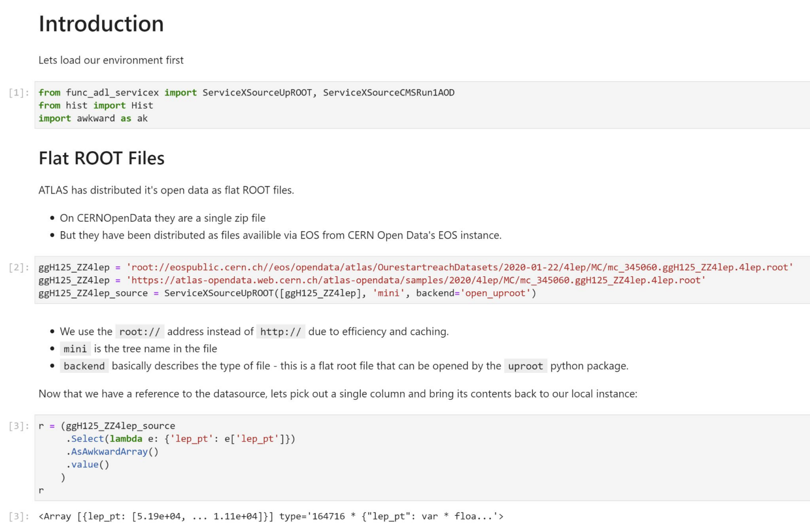
In the end, many parts of coffea should disappear into dedicated packages… until then, missing gaps in functionality have a home here.

documentation

Notebooks

We have 3 notebooks to work through!

- Introduction to func_adl, ServiceX, and coffea
- The ATLAS Higgs Discovery Dataset
- The CMS Higgs Discovery Dataset



Introduction

Lets load our environment first

```
[1]: from func_adl_servicex import ServiceXSourceUpROOT, ServiceXSourceCMSRun1AOD
     from hist import Hist
     import awkward as ak
```

Flat ROOT Files

ATLAS has distributed it's open data as flat ROOT files.

- On CERNOpenData they are a single zip file
- But they have been distributed as files availible via EOS from CERN Open Data's EOS instance.

```
[2]: ggH125_ZZ4lep = 'root://eospublic.cern.ch//eos/opendata/atlas/OurestartreachDatasets/2020-01-22/4lep/MC/mc_345060.ggH125_ZZ4lep.4lep.root'
     ggH125_ZZ4lep = 'https://atlas-opendata.web.cern.ch/atlas-opendata/samples/2020/4lep/MC/mc_345060.ggH125_ZZ4lep.4lep.root'
     ggH125_ZZ4lep_source = ServiceXSourceUpROOT([ggH125_ZZ4lep], 'mini', backend='open_uproot')
```

- We use the `root://` address instead of `http://` due to efficiency and caching.
- `mini` is the tree name in the file
- `backend` basically describes the type of file - this is a flat root file that can be opened by the `uproot` python package.

Now that we have a reference to the datasource, lets pick out a single column and bring its contents back to our local instance:

```
[3]: r = (ggH125_ZZ4lep_source
          .Select(lambda e: {'lep_pt': e['lep_pt']})
          .AsAwkwardArray()
          .value()
         )
     r

[3]: <Array [{lep_pt: [5.19e+04, ... 1.11e+04]}] type='164716 * {"lep_pt": var * floa...'>
```

Tutorial GitHub Repo

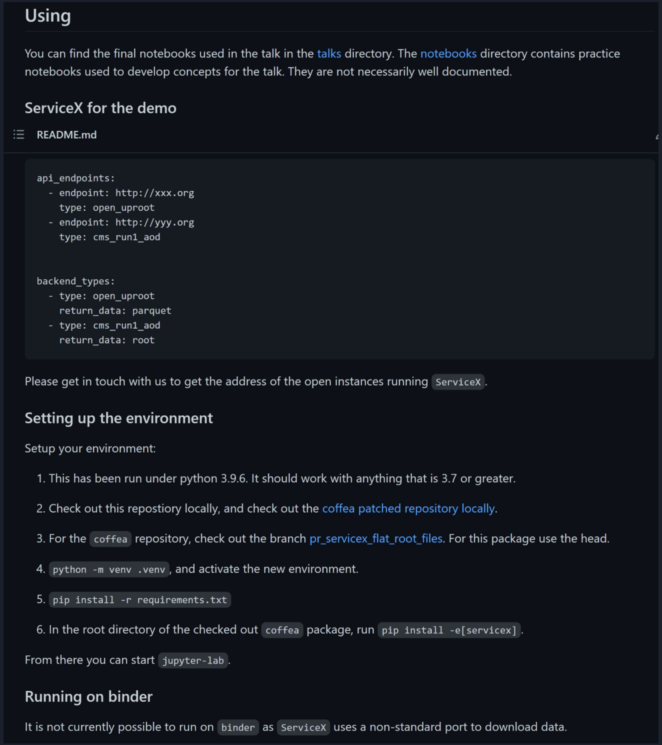
Team Effort

- Data and MC: ATLAS & CMS Collaborations for all the data, and making it public!
- ServiceX: Ben Galewsky and Andrew Eckart and Suchandra Thapa and everyone else on the team
- CMS Run 1 AOD Transformer: Baidyanath Kundu
- Uproot/Flat ROOT backend: Mason Proffit
- IRIS-HEP for supporting a large fraction of these people in one way or another.
- CMS Awkward Code: Brian Cruz

Running This Yourself...

Tutorial GitHub Repo

- Instructions are on the repo
- We are happy to give out the ServiceX end-points (unofficial atm)
- Some integration with coffea is in beta
- Not yet obvious that CERN Open Data portal can handle the data load!
- Binder is not supported



Using

You can find the final notebooks used in the talk in the talks directory. The notebooks directory contains practice notebooks used to develop concepts for the talk. They are not necessarily well documented.

ServiceX for the demo

```
api_endpoints:
  - endpoint: http://xxx.org
    type: open_uproot
  - endpoint: http://yyy.org
    type: cms_run1_aod


backend_types:
  - type: open_uproot
    return_data: parquet
  - type: cms_run1_aod
    return_data: root
```

Please get in touch with us to get the address of the open instances running `ServiceX`.

Setting up the environment

Setup your environment:

1. This has been run under python 3.9.6. It should work with anything that is 3.7 or greater.
2. Check out this repostiory locally, and check out the coffea patched repository locally.
3. For the `coffea` repository, check out the branch pr_servicex_flat_root_files. For this package use the head.
4. `python -m venv .venv`, and activate the new environment.
5. `pip install -r requirements.txt`
6. In the root directory of the checked out `coffea` package, run `pip install -e[servicex]`.

From there you can start `jupyter-lab`.

Running on binder

It is not currently possible to run on `binder` as `ServiceX` uses a non-standard port to download data.