Van deeltjes naar data

Exabyte-scale dataprocessing bij CERN

Florine de Geus (florine.de.geus@cern.ch)

26-09-2024

Wie ben ik?

👩‍💻 Florine de Geus

🏫 Sinds januari: PhD-student bij CERN & Universiteit Twente

→ Hiervoor al een jaar als *technical student* mogen rondlopen

Middelbare school VWO met NT-profiel

Bachelor Informatica, UvA

Master Software Engineering, UvA

Hoe ben ik bij CERN terechtgekomen?

Hoe het allemaal begon, februari 2022...

Hi Florine,

How are you? How is the SE program?

I am writing to ask whether you are interested in a project at CERN for your graduation (and potentially after). Let me know - and we can discuss in a short meeting.

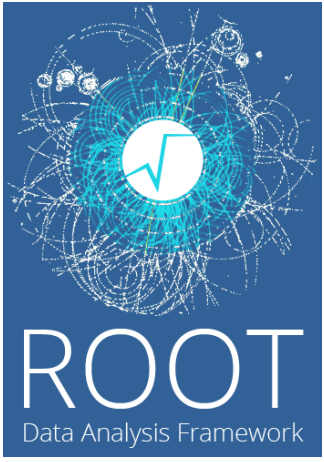
Regards,

Wat doe ik bij CERN?

ROOT: Software voor het opslaan en verwerken van natuurkundedata

(Bijna) alle ingrediënten om van ruwe data tot nieuwe natuurkundige inzichten te komen:

- Schrijven en lezen van data (I/O)
- Definiëren van analysestappen
- Vullen van histogrammen
- Fitten van data
- Statistische analyse
- Genereren van plots



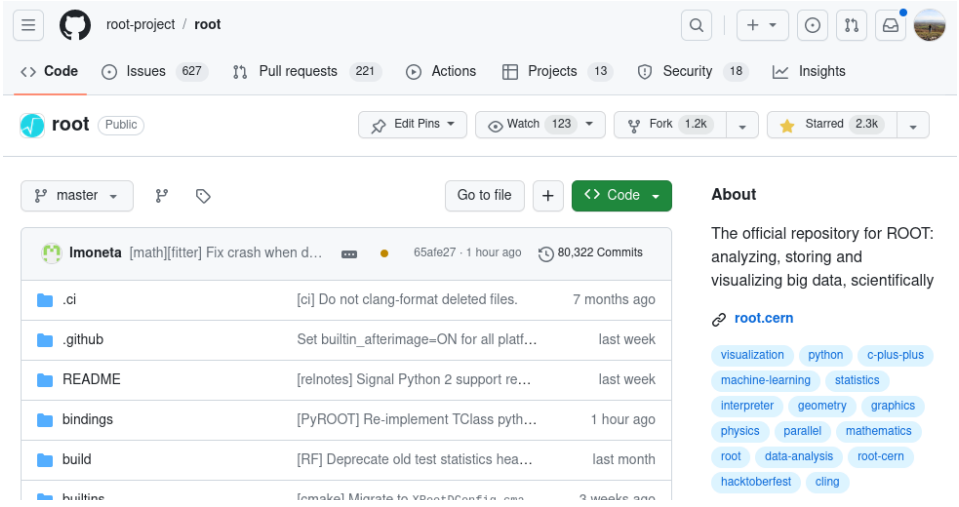
Wat doe ik bij CERN?

ROOT: Software voor het opslaan en verwerken van natuurkundedata

(Bijna) alle ingrediënten om van ruwe data tot nieuwe natuurkundige inzichten te komen:

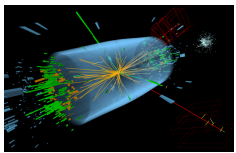
- **Schrijven en lezen van data (I/O)**
- Definiëren van analysestappen
- Vullen van histogrammen
- Fitten van data
- Statistische analyse
- Genereren van plots

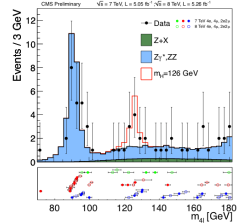
ROOT is Open Source!



Van deeltjes naar data

Van deeltjes naar data

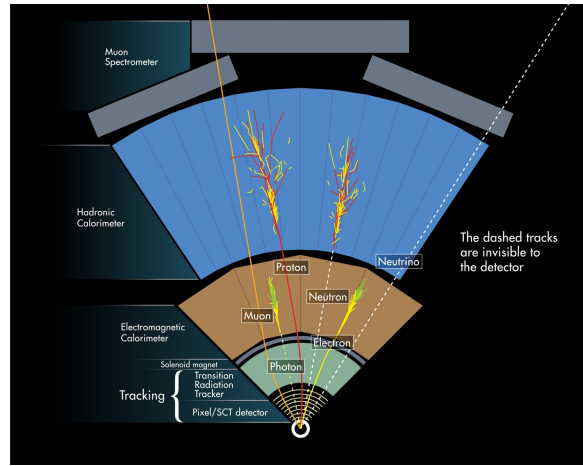




Gebaseerd op Induction to Data Processing of LHC Experiments door Danilo Piparo

Ruwe data

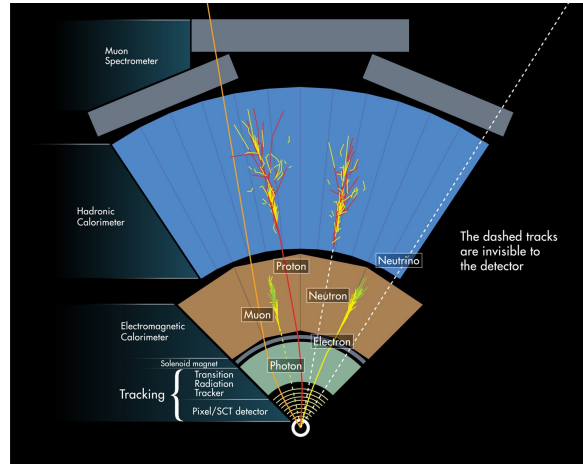
- Één *event* (botsing) = 1-1.5 MB
- Per seconde: 40 miljoen (!) events
- *Trigger* filtert potentieel interessante botsingen
  - Combinatie tussen speciale hardware, FPGAs en GPUs
- 1000 events per seconde blijven over voor reconstructie



Bron: CERN-EX-1301009

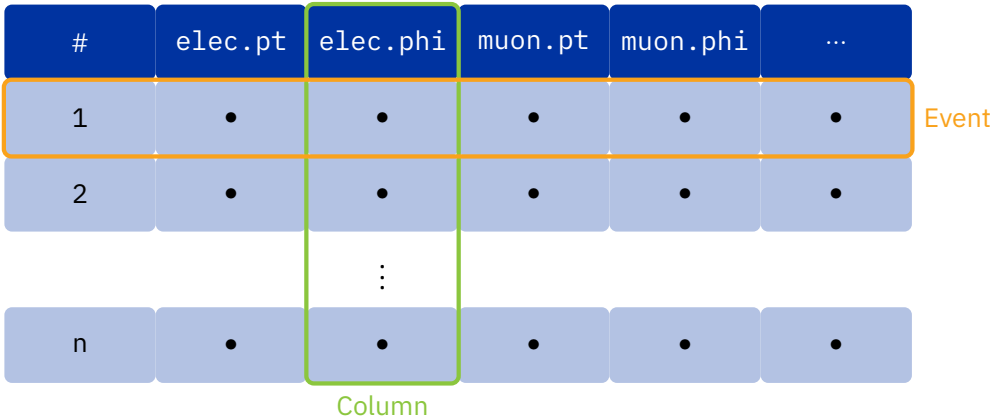
Reconstructie

- 🧀 maken van de detectordata
- *Tracking*: reconstructie van het pad dat een deeltje aflegt
- Onderscheiden van elektronen, muonen, fotonen, *jets*, ...
  - En hun gemeten eigenschappen
- Vanaf hier wordt ROOT relevant!



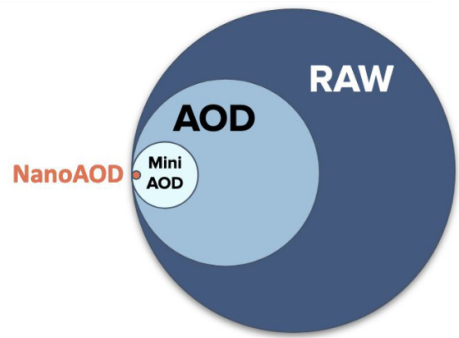
Bron: CERN-EX-1301009

Hoe worden deeltjes opgeslagen?



Na de reconstructie

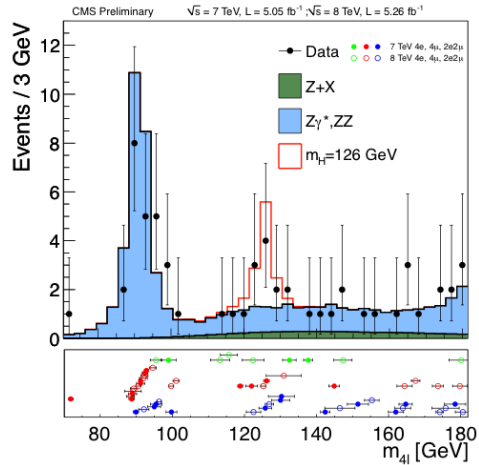
- *AOD = Analysis Object Data*
- (Meestal) verdere verwerking voor ze bij de natuurkundigen terecht komen:
  - Toevoegen/verwijderen van kolommen
  - Corrigeren/kalibreren van data
- Kleinere dataset bij elke iteratie
  - Maar: elke tussenstap moet worden bewaard



Bron: Induction to Data Processing of LHC Experiments

Analyse

- *Event loop* waarin data voor de laatste keer gefilterd en verwerkt wordt...
  - Analyse-specifiek
  - Bv: bereken de invariante muonmassa, maar alleen voor events met > 2 muonen
- Vullen van histogrammen
  - Manier om te zien of en hoe de geobserveerde data afwijkt van status-quo
  - Kan een nieuw deeltje betekenen!



Bron: CMS-PHO-EVENTS-2013-004-1

De data-uitdaging

1 EB
=
1 000 000 000 GB

- Er is tot nu toe meer dan 2 *exabyte* aan LHC data, opgeslagen in datacentra over de hele wereld
- Jaarlijks wordt 1 *exabyte* ‘aangeraakt’ (i.e. geschreven of gelezen)[1]
- Het overgrote deel hiervan is in ROOT’s huidige dataformat: **TTree**

[1] Bron: Key Facts and Figures – CERN Data Centre

De data-uitdaging

1 EB
=
1 000 000 000 GB

- Er is tot nu toe meer dan 2 *exabyte* aan LHC data, opgeslagen in datacentra over de hele wereld
- Jaarlijks wordt 1 *exabyte* 'aangeraakt' (i.e. geschreven of gelezen)[1]
- Het overgrote deel hiervan is in ROOT's huidige dataformat: **TTree**

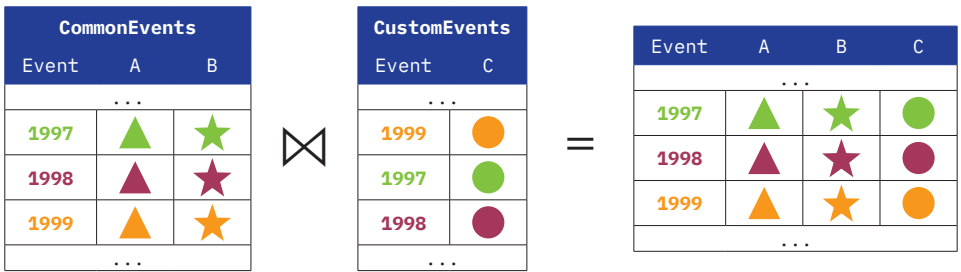
Vanaf 2029 (LHC Run 4) wordt **10x zo veel data**[1] verwacht!

- Het ROOT-team is druk aan het werk aan de opvolger: **RNTuple**
  - Hiermee kunnen we (momenteel) data ~25% efficiënter opslaan en vanaf 2-3x sneller verwerken
  - Doel is om dit nog verder te optimaliseren, o.a. als onderwerp van mijn PhD

[1] Bron: Key Facts and Figures – CERN Data Centre

Mijn onderzoek in een notendop

Hoe kunnen we verschillende datasets (efficient) combineren?



Simpel voor drie rijen, maar wat als we er drie miljard hebben?

Tot slot

Persoonlijke ervaringen en lessen

- De hoofdtaak van CERN is natuurlijke deeltjesfysica – veel meer andere disciplines en domeinen nodig om dit voor elkaar te krijgen!

Persoonlijke ervaringen en lessen

- De hoofdtaak van CERN is natuurlijke deeltjesfysica – veel meer andere disciplines en domeinen nodig om dit voor elkaar te krijgen!
- Heel cool om te realiseren dat waar jij aan werkt bij gaat dragen aan een volgende (misschien wel grote) ontdekking, ook al is het maar een klein onderdeel

Persoonlijke ervaringen en lessen

- De hoofdtaak van CERN is natuurlijke deeltjesfysica – veel meer andere disciplines en domeinen nodig om dit voor elkaar te krijgen!
- Heel cool om te realiseren dat waar jij aan werkt bij gaat dragen aan een volgende (misschien wel grote) ontdekking, ook al is het maar een klein onderdeel
- Je ontmoet ontzettend veel mensen!

Persoonlijke ervaringen en lessen

- De hoofdtaak van CERN is natuurlijke deeltjesfysica – veel meer andere disciplines en domeinen nodig om dit voor elkaar te krijgen!
- Heel cool om te realiseren dat waar jij aan werkt bij gaat dragen aan een volgende (misschien wel grote) ontdekking, ook al is het maar een klein onderdeel
- Je ontmoet ontzettend veel mensen!
- Voor iemand die wel altijd geïntereseerd was in natuurkunde maar er niet bijzonder goed in was op school is dit een ideale positie :-)

Mogelijkheden binnen CERN voor studenten

- Summer student programme
  - 8-13 weken
  - 3 jaar studie-ervaring nodig
- Technical student programme
  - 4-12 maanden
  - 1,5 jaar studie-ervaring nodig
- Short-term internship programme
  - 1-6 maanden

Bron: CERN Document Server

Veel plezier tijdens de rest van jullie bezoek!