CERN openlab

Big data analysis and machine learning in the cloud

CERN openlab Summer Student Programme

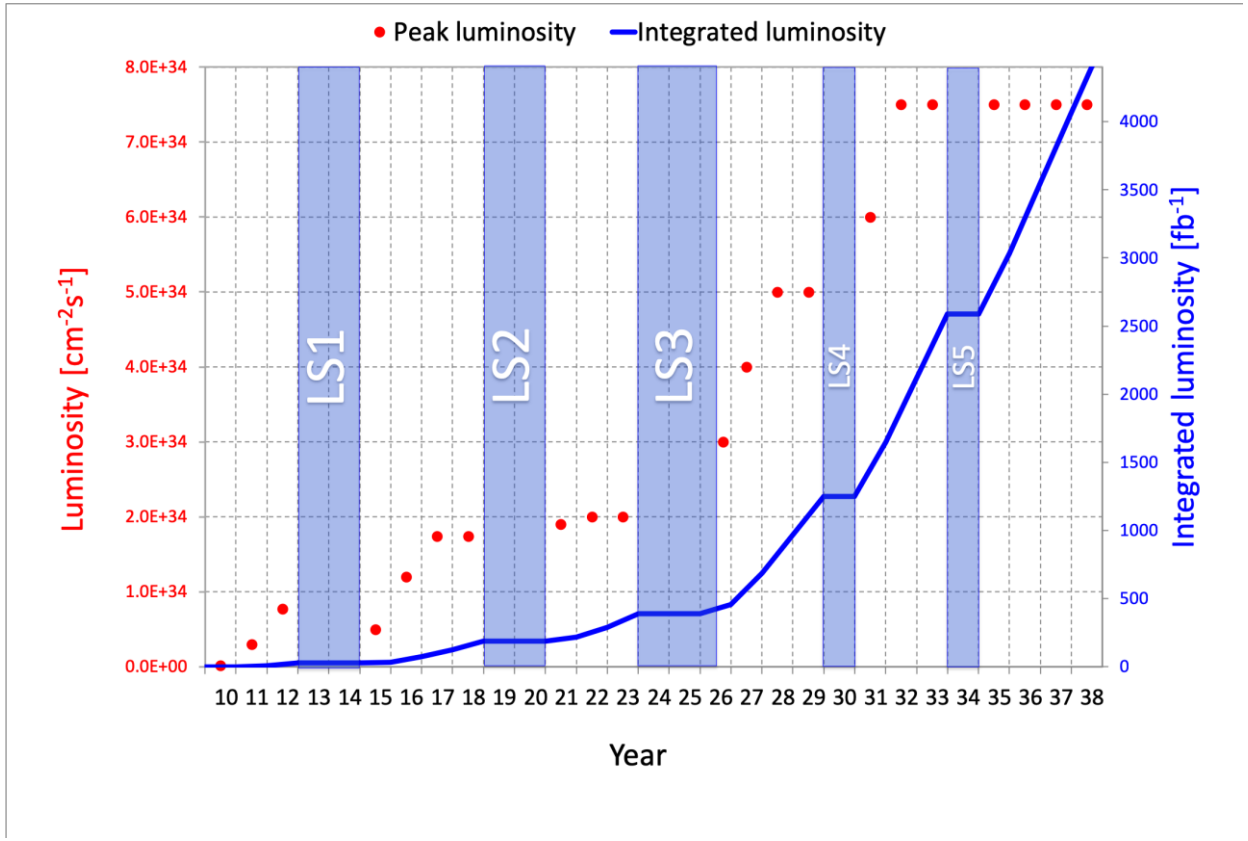
Michał Bień

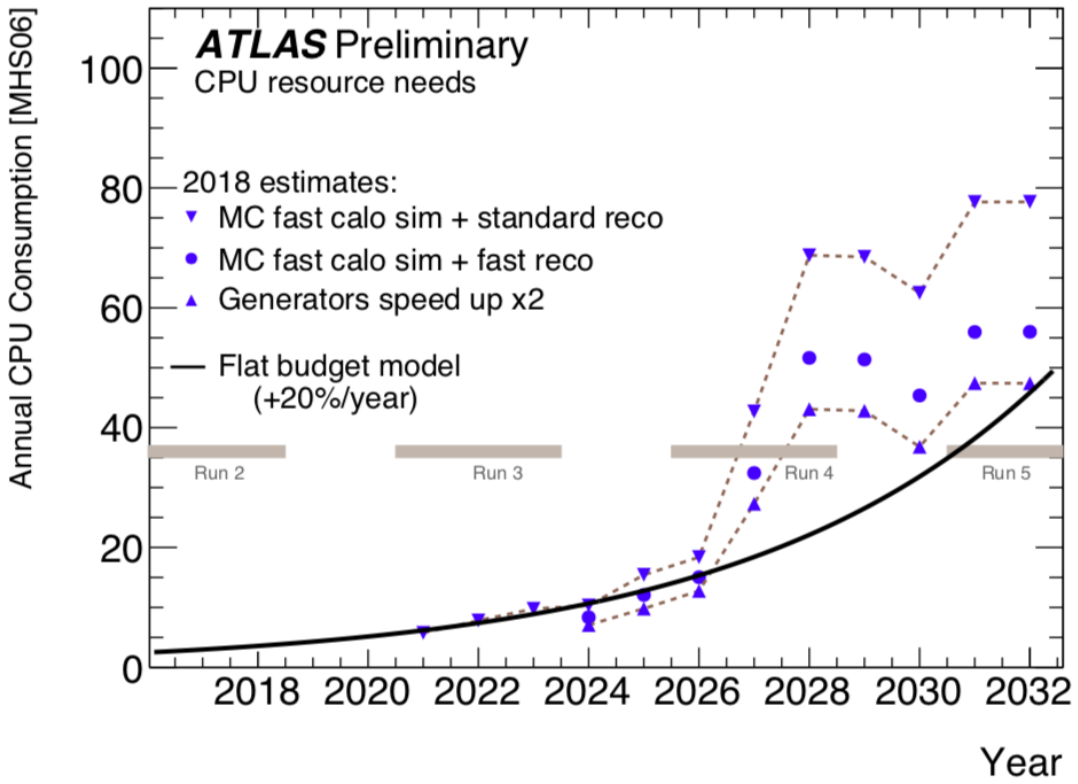
Supervisors: Riccardo Castellotti, Luca Canali

15.08.2019

Motivation

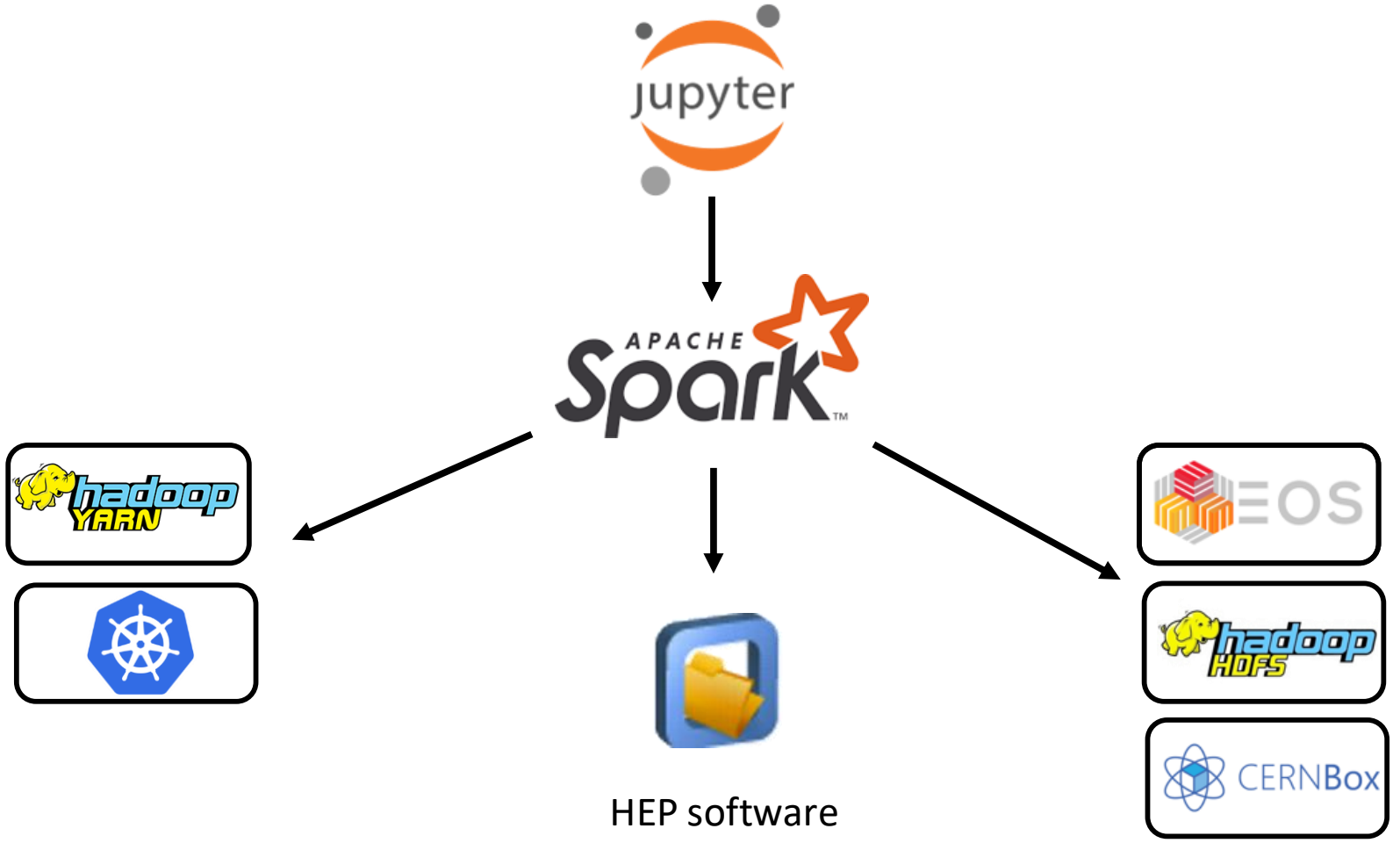
High luminosity LHC and its data production rate



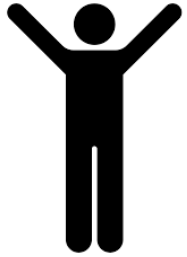


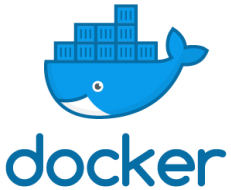
Credits: https://lhc-commissioning.web.cern.ch, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults

Data analytics solutions



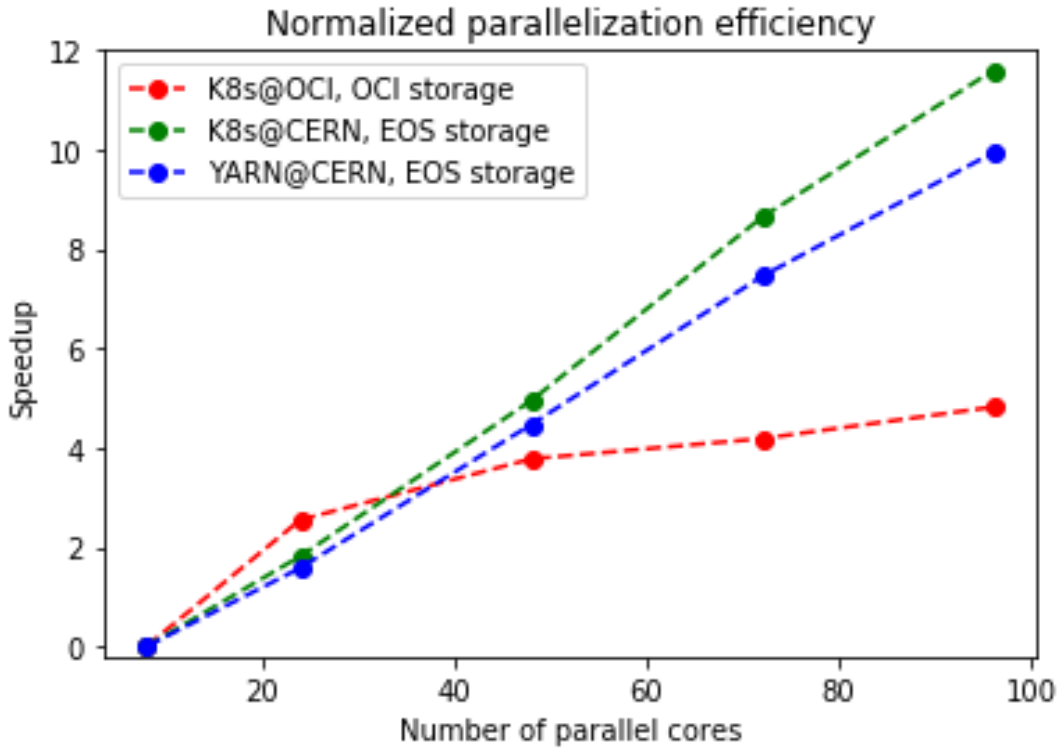
Data analytics in the cloud



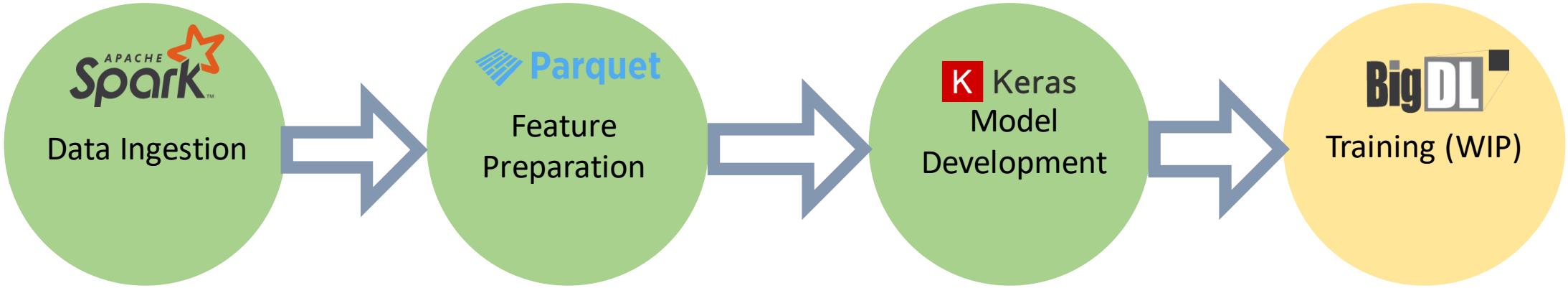


CMS Big Data Reduction:

- Highly memory- and IO-intensive
- 22TB of data to reduce



Spark DL Trigger



- Multi-step interactive pipeline
- Deep learning training with Intel BigDL
- Highly CPU intensive
- Highly communication-intensive
- Possible further enhancements

In the meantime...

OCI-HDFS-Connector:
1 bug fixed
1 issue reported

EPEL Repositories:
1 issue reported

2 Helm charts prepared

spark-root & root4j
Fixed bugs in both
Triggered new version release

Apache Spark:
New instrumentation to measure OCI metrics

Countless Docker images built and deployed

Further steps

- Distributed TensorFlow
- Exact, online cost measurement
- Possible integration of public cloud and CERN services
- Further OCI scalability tests
- Training on GPUs!

Are we satisfied?

Thank you for your attention!

And HUGE "Thank you!" to:
Riccardo and Luca
IT-DB-SAS team
CERN openlab

michal.bien@cern.ch