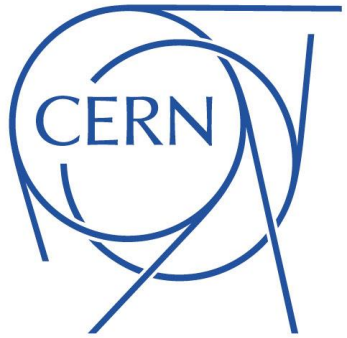


Tools, Development and Selected Projects

J.P. Morgan - CERN Big Data Meeting

Evangelos Motesnitsalis

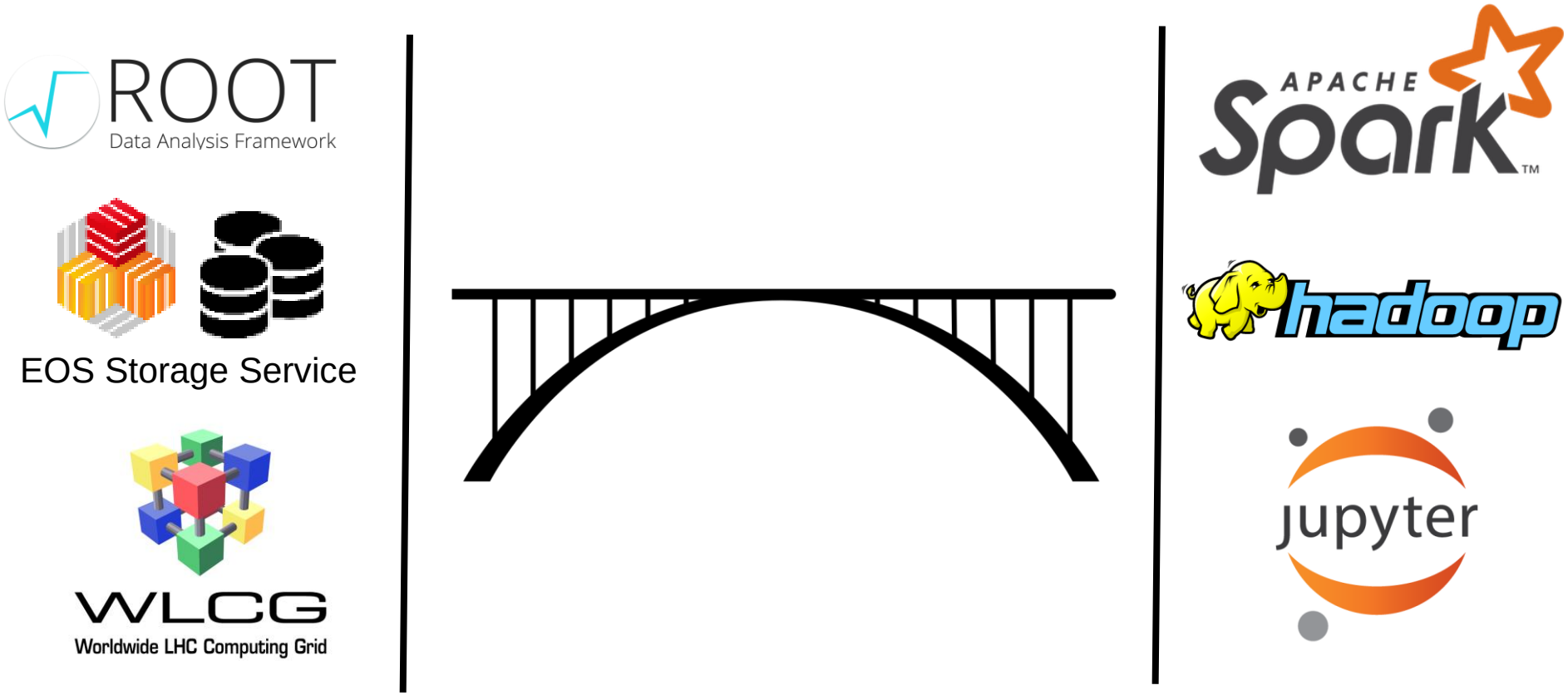
07/05/2018

CERN openlab

Tools and Development

Connecting the Old and the New

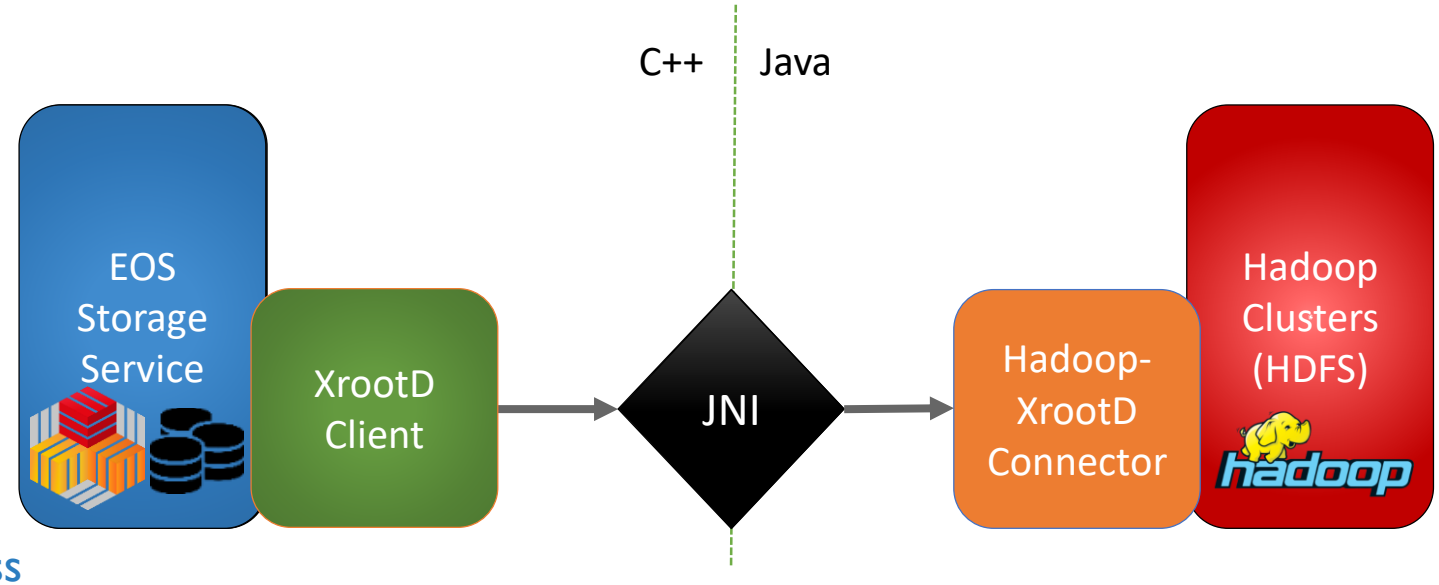
Problem Definition



Hadoop – XRootD Connector

Connecting XrootD-based Storage Systems with Hadoop and Spark

- A Java library that connects to the XrootD client via JNI
- It reads files from the EOS Storage Service directly
- Slower Read – Broader Data Access



https://github.com/cerndb/hadoop-xrootd

Spark - Root

A Scala library which implements DataSource for Apache Spark

Spark can read ROOT TTrees and infer their schema

Root files are imported to Spark Dataframes/Datasets/RDDs

Developed by DIANA-HEP

https://github.com/diana-hep/spark-root/

Future Plans

Spark on Kubernetes Service

Spark on Kubernetes Service

Leveraging the Kubernetes support in Spark 2.3

- Prototype of Spark on Kubernetes over OpenStack and built spark images and tooling
- Under active development
- Work on the **cern-spark-service** python package [Early Alpha Release!]

https://pypi.python.org/pypi/cern-spark-service

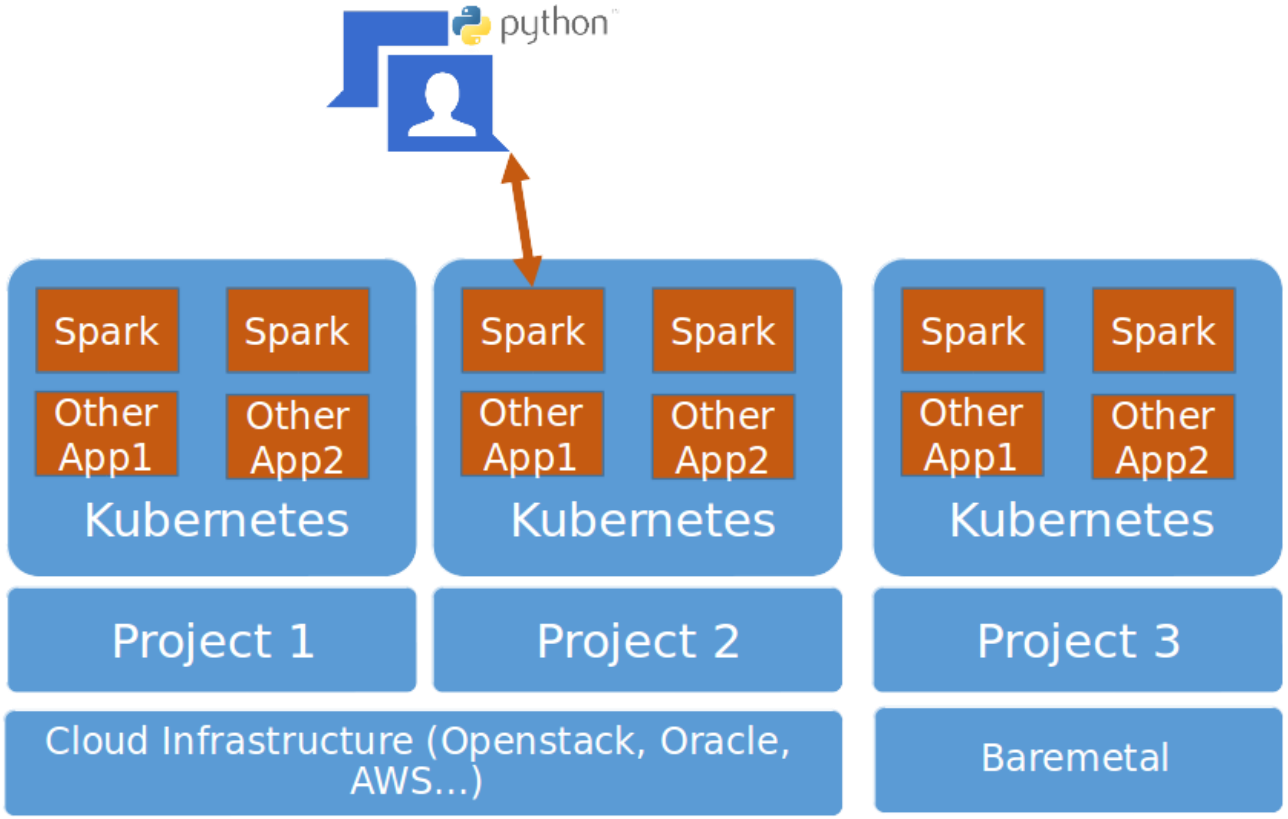
Spark on Kubernetes Service

Leveraging the Kubernetes support in Spark 2.3

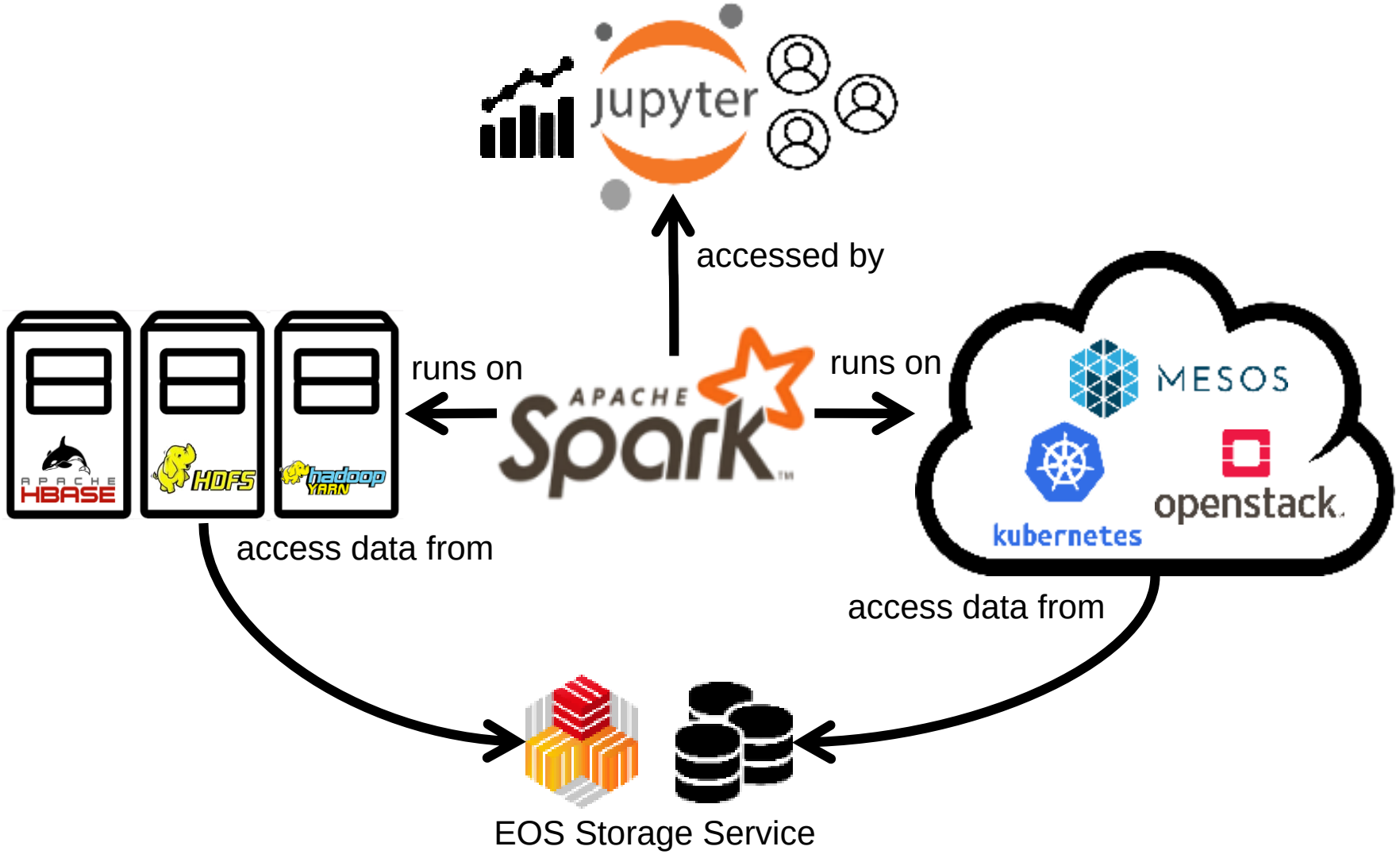
1. Create Kubernetes cluster and initialize its dependencies

2. Submit Spark jobs to Spark using **cern-spark-submit** python package

3. Our docker images will deploy and run the Spark application over Kubernetes as if it was on Hadoop/YARN



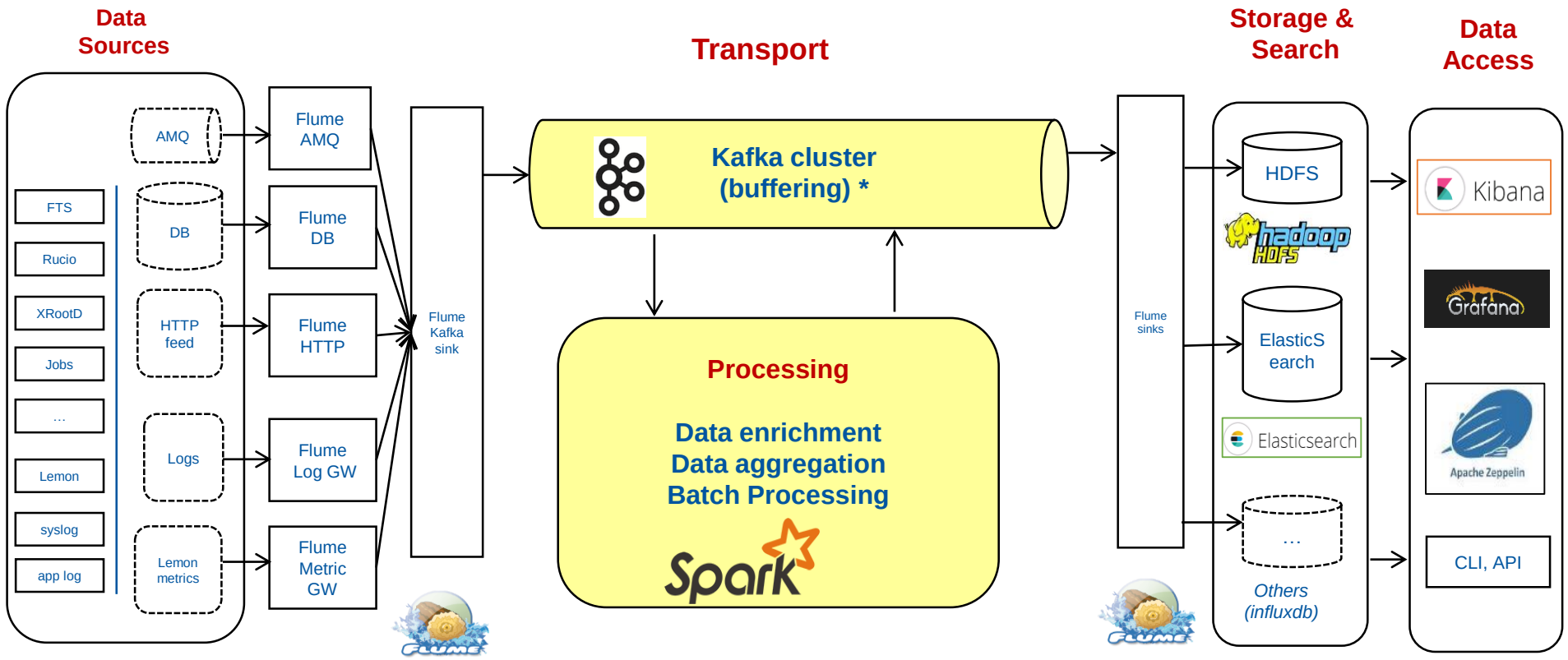
Analytics Platform



Selected Projects and Platforms

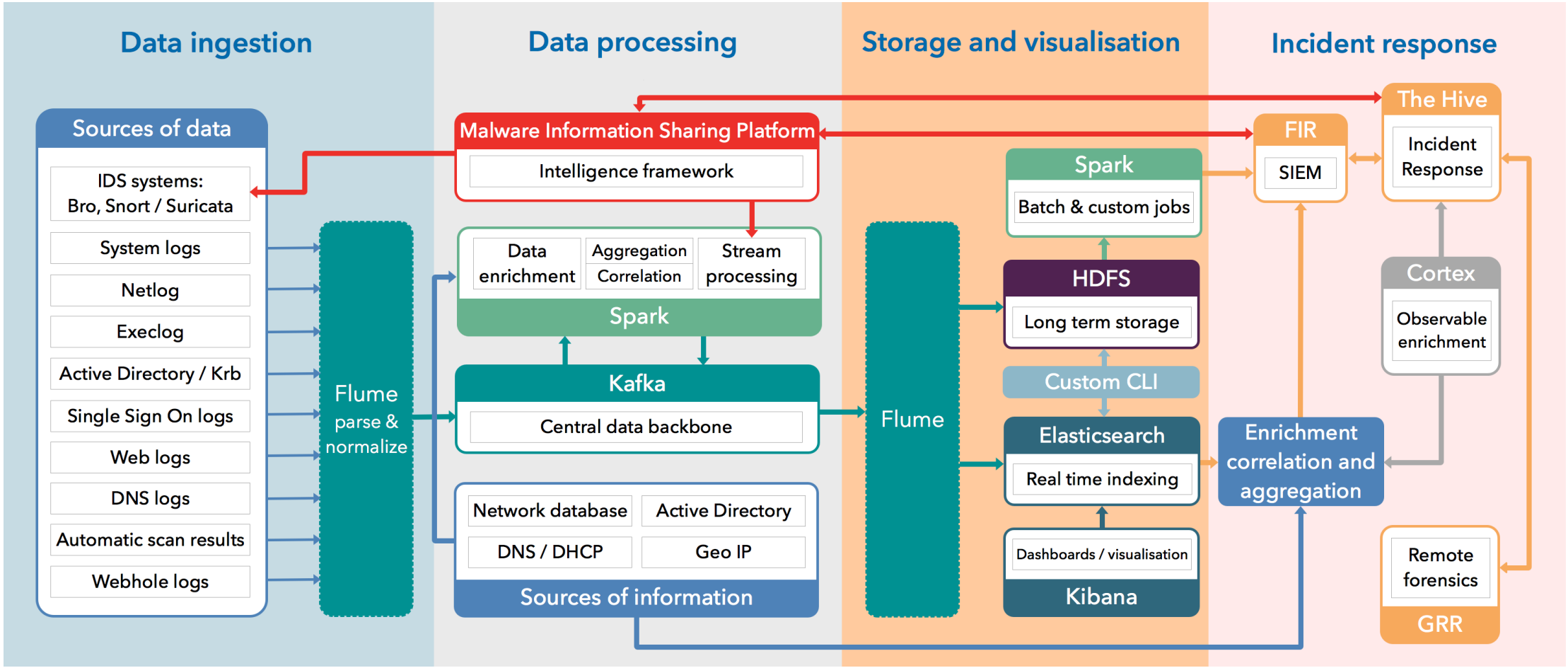
Data Center and WLCG Monitoring Systems

- Critical for Data Center operations and WLCG
- 200M events/day
  500 GB/day
- Proved effective in several occasions



Credits: IT-CM-MM

Computer Security Intrusion Detection



Credits: CERN security team, IT-DI

CMS Data Reduction Facility

Performing Physics Analysis and Data Reduction with Apache Spark

Investigate new ways to analyse physics data and improve resource utilization and time-to-physics

Until today, high energy physics analysis is done with the ROOT Framework

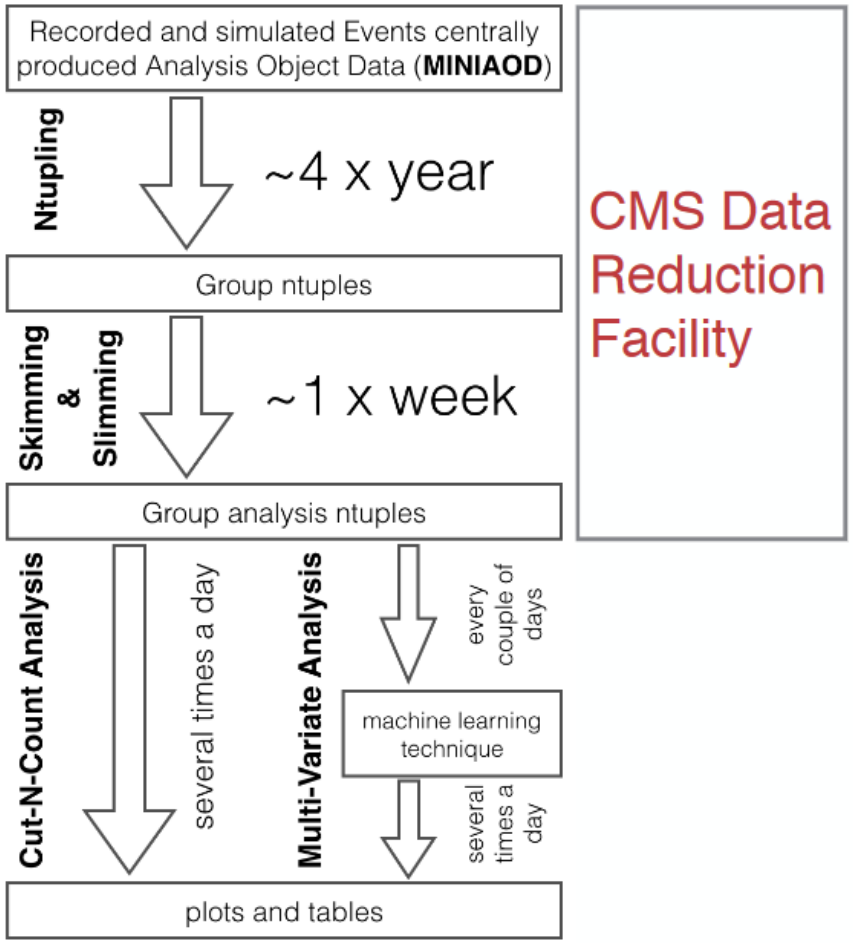
We now have fully functioning Analysis and Reduction examples tested over CMS Open Data (1 TB)

We started scaling – goal is 1 PB

Root files are imported with « spark-root »

Files are accessed from the EOS Storage Service with the « Hadoop-XRootD Connector »

CMS Data Reduction Facility



- Produce reduced data based on potential complicated user queries
- If sucessful, this type of facility could be a big shift for High Energy Physics
- Make High Energy Physics more open to the Big Data community

Thank you