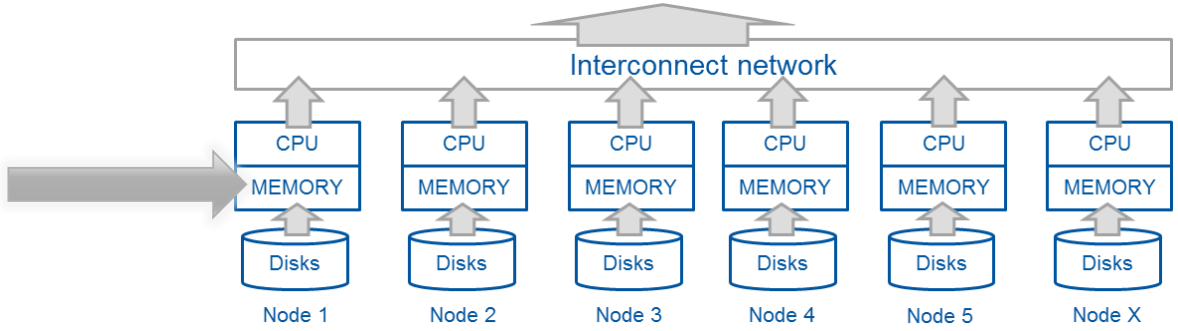
Evolution of Hadoop and Spark service at CERN

Zbigniew Baranowski
CERN IT-DB Hadoop and Spark Service, Streaming Service
May 17th, 2018

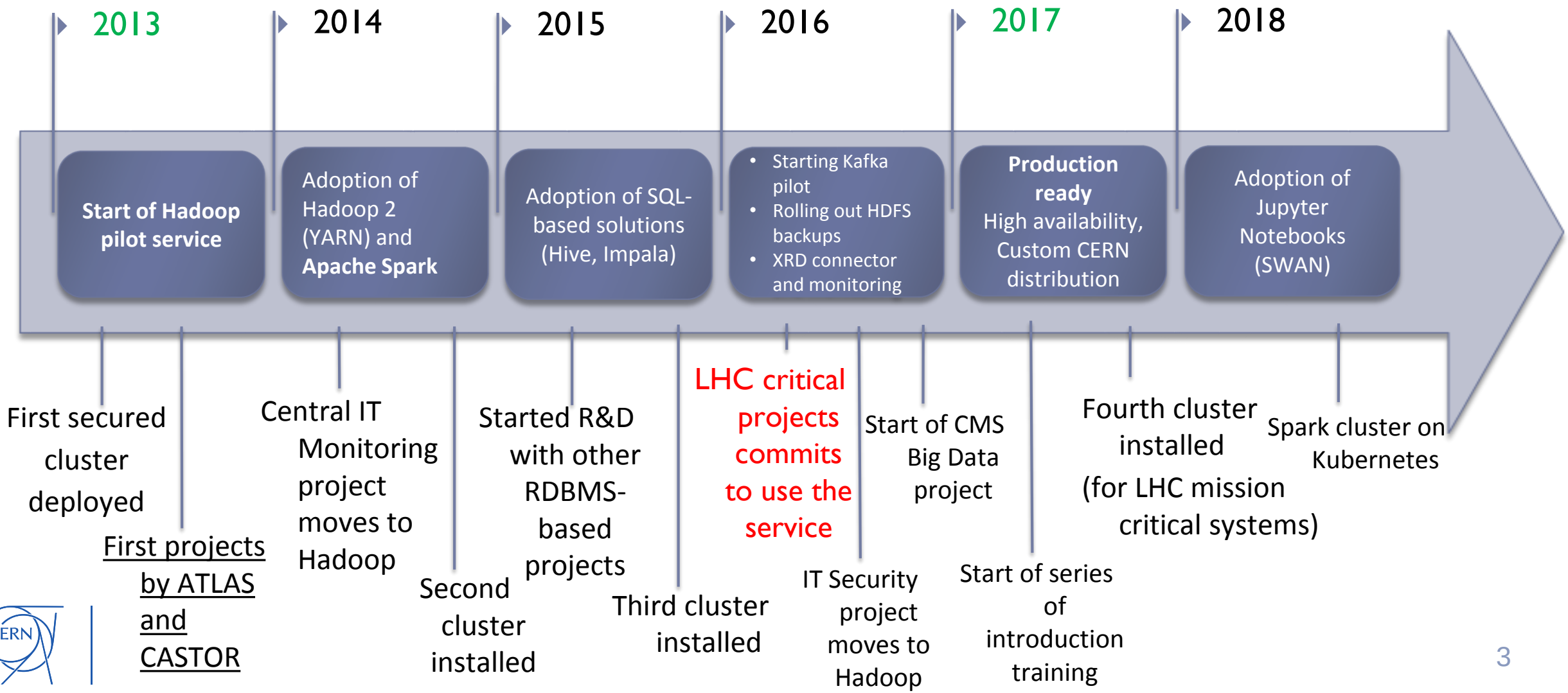
Hadoop and Spark for big data analytics

- Distributed systems for data processing
  - Storage and multiple data processing interfaces
  - Can operate at scale by design (shared nothing)
  - Typically on clusters of commodity-type servers/cloud
  - Many solutions target data analytics and data warehousing
  - Can do much more: stream processing, machine learning
- Already well established in the industry and open source

Scale-out data processing



CERN Hadoop Service - Timeline



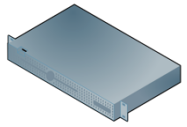
Hadoop Service at CERN IT

- Setup and run the infrastructure
- Support user community
  - Provide consultancy
  - Ensure knowledge sharing
  - Train on the technologies

Hadoop service in numbers

6 clusters

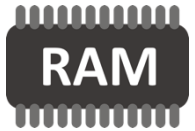
- 4 production (bare-metal)
- 2 QA clusters (VMs)



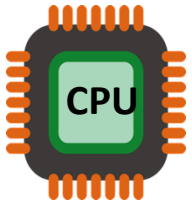
110+ physical servers

40+ virtual machines

14+ PBs of Storage

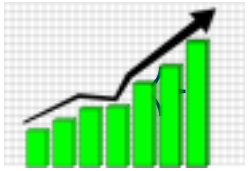


20+ TB of Memory



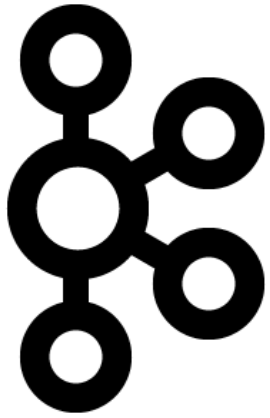
1500+ physical cores

HDDs and SDDs

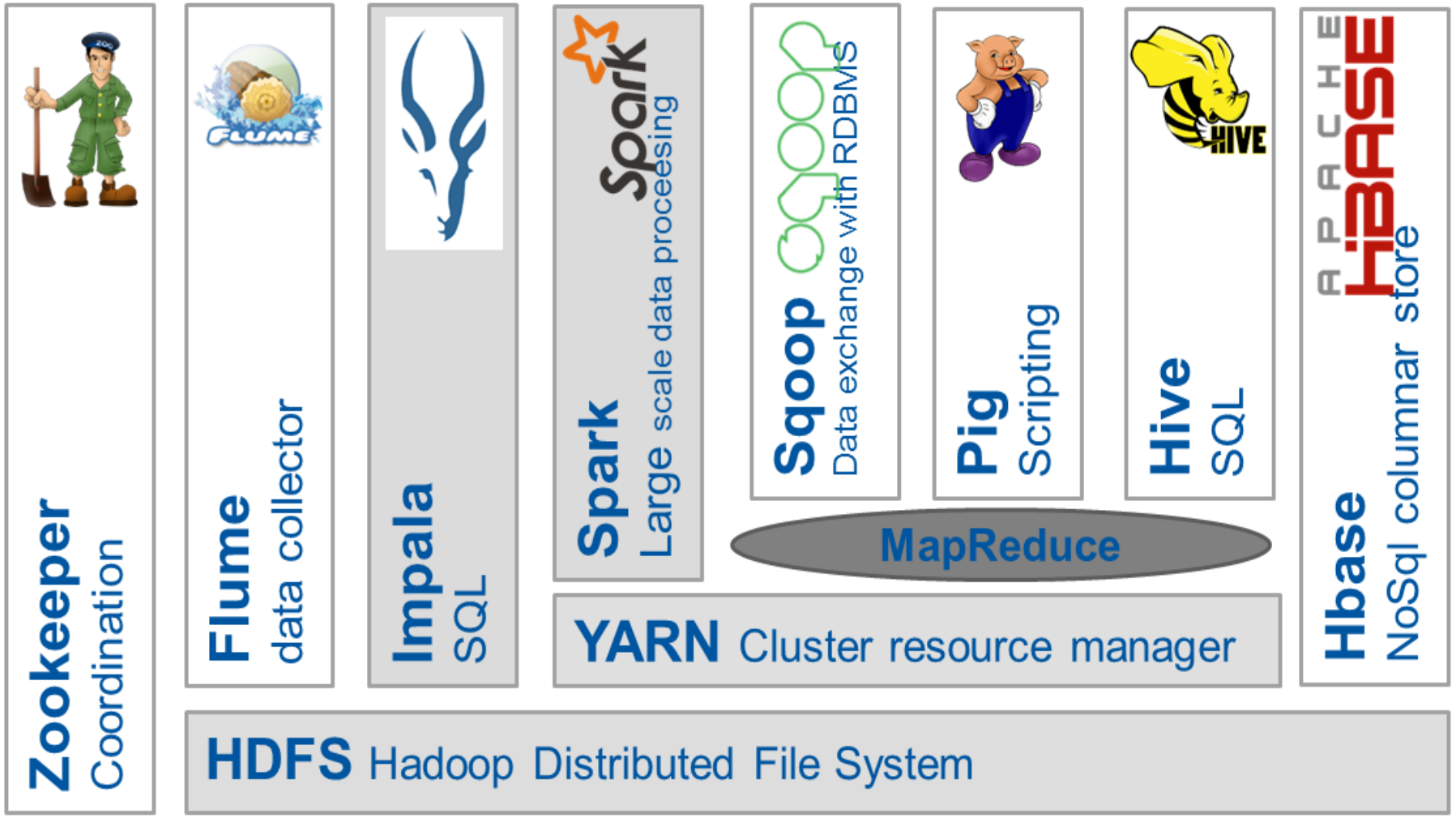


Data growth: 4 TB per day

Overview of available components in 2018



Kafka:
streaming
and ingestion



Hadoop and Spark production deployment

- Software distribution
  - Cloudera (since 2013)
  - Vanilla Apache (since 2017)

- Installation and configuration
  - CentOS 7.4
  - custom Puppet module

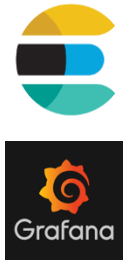
- Security
  - authentication Kerberos
  - fine-grained authorization integrated with e-groups (since 2018)

- High availability (since 2017)
  - automatic master failover
  - for HDFS, YARN and HBASE

- Rolling change deployment
  - no service downtime
  - transparent in most of the cases

- Host monitoring and alerting
  - via CERN IT Monitoring infrastructure

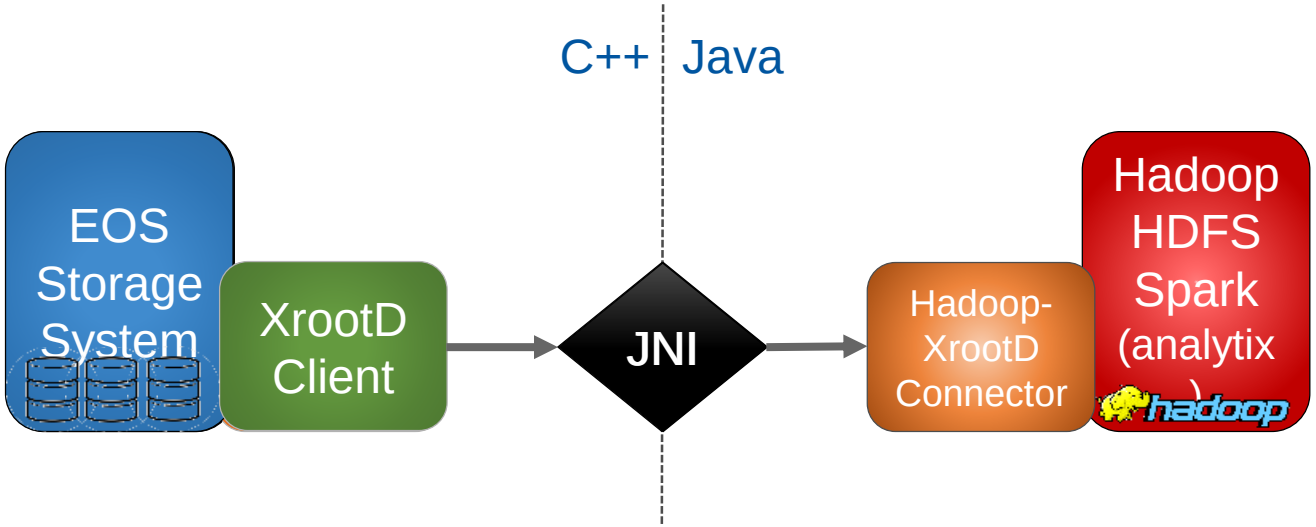
- Service level monitoring (since 2017)
  - metrics integrated with: Elastic + Grafana
  - custom scripts for availability and alerting



- HDFS backups (since 2016)
  - Daily incremental snapshots
  - Sent to tapes (CASTOR)

XRootD connector for Hadoop and Spark

- A library that binds Hadoop-based file system API with XRootD native client
  - Developed by CERN IT
- Allows most of components from Hadoop stack (Spark, MapReduce, Hive etc) to read from/write to EOS and CASTOR directly
  - Works with Grid certificates and Kerberos for authentication
  - Used for: HDFS backups, performing analytics on data stored on EOS (CERNBox)



Moving to Apache Hadoop distribution (since 2017)

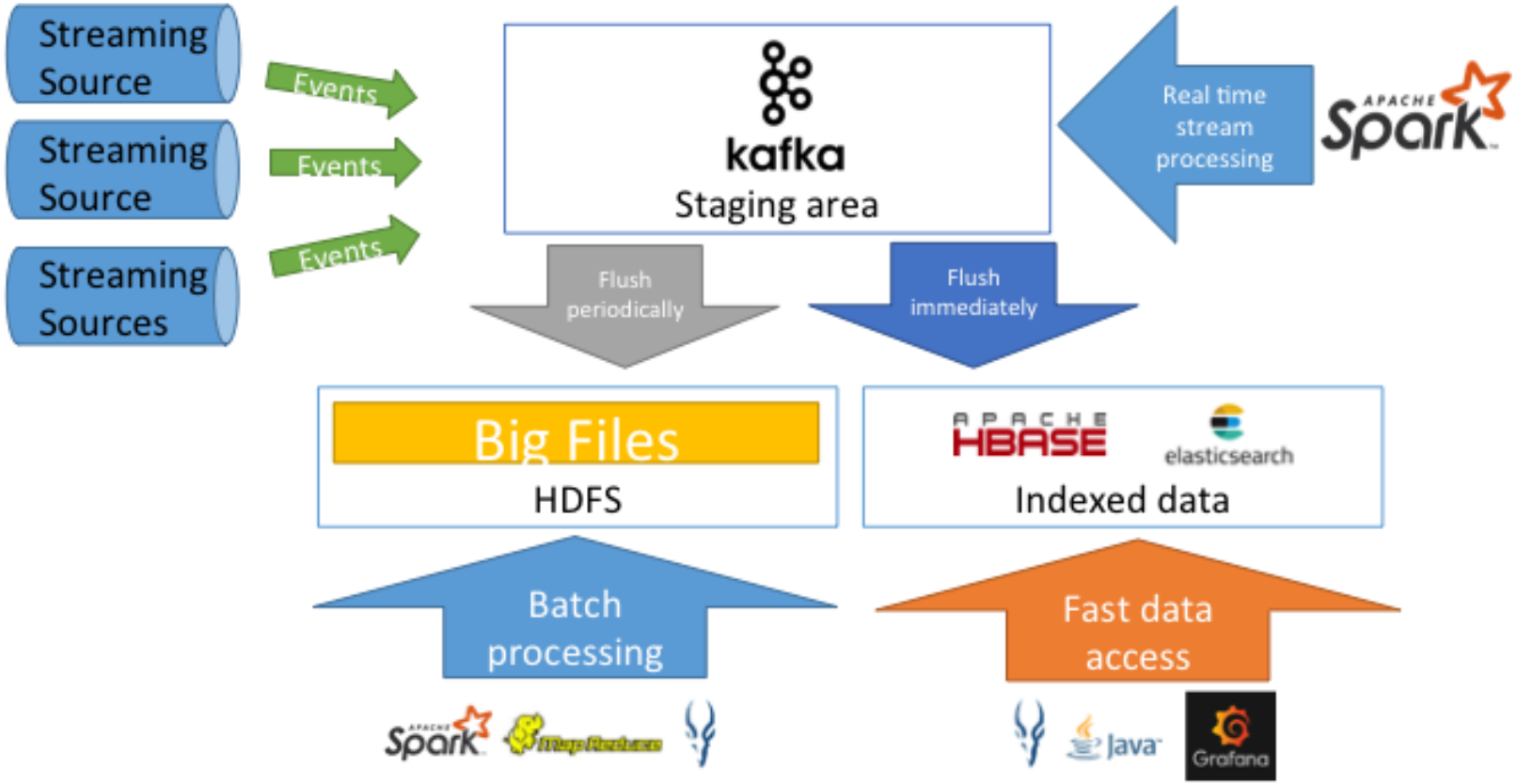
- Better control of the core software stack
  - Independent from a vendor/distributor
  - In-house compilation
  - Enabling non-default features (compression algorithms, R for Spark)
  - Adding critical patches (that are not ported in upstream)

- We do rpm packaging for core components
  - HDFS and YARN, Spark, HBase

- Streamlined development
  - Available on Maven Central Repository

Apache Kafka - data streaming at scale

- Apache Kafka streaming platform becomes a standard component for modern scalable architectures
- Started providing Kafka as a pilot service in 2017



SWAN – Jupyter Notebooks On Demand

- Service for web based analysis (SWAN)
  - Developed at CERN, initially for physics analysis by EP-SFT and IT-ST
- An interactive platform that combines code, equations, text and visualizations
  - Ideal for exploration, reproducibility, collaboration
- Fully integrated with Spark and Hadoop at CERN (2018)
  - Python on Spark (PySpark) at scale
  - Modern, powerful and scalable platform for data analysis
  - Web-based: no need to install any software

FILE EDIT VIEW INSERT CELL KERNEL HELP Trusted Python 2

Code

Do the heavylifting in spark and collect aggregated view to panda DF

In [11]:
```
df_loadAvg_pandas = spark.sql("SELECT submitter_host, \
                                      avg(body.LoadAvg) as avg, \
                                      hour(from_unixtime(timestamp / 1000, 'yyyy-MM-dd HH:mm:ss')) as hr \
                               FROM loadAvg \
                               WHERE submitter_hostgroup = 'hadoop/itdb/datanode' \
                               AND dayofmonth(from_unixtime(timestamp / 1000, 'yyyy-MM-dd HH:mm:ss')) = 15 \
                               GROUP BY hour(from_unixtime(timestamp / 1000, 'yyyy-MM-dd HH:mm:ss')), submitter_host")\
                    .toPandas()
```

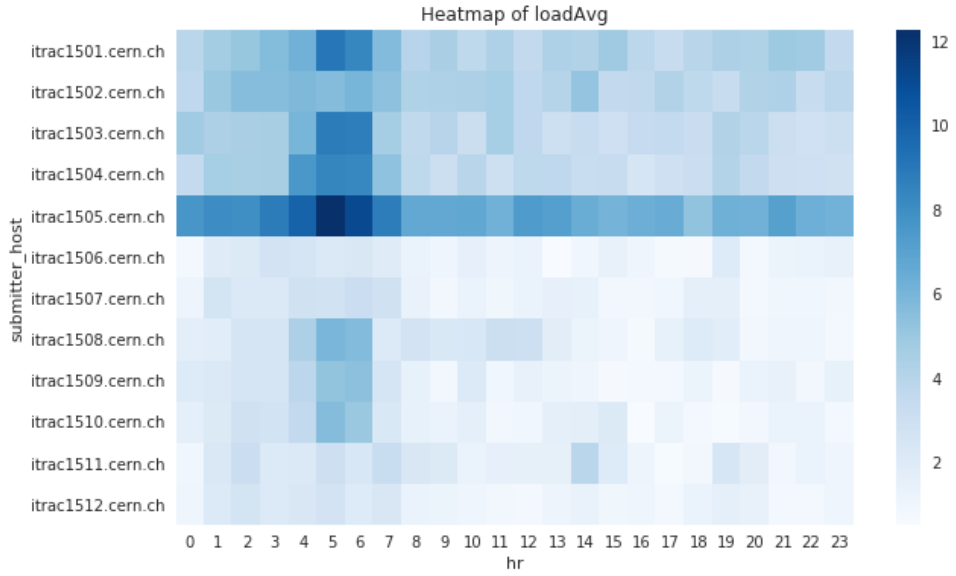
Apache Spark: 90 EXECUTORS 180 CORES Jobs: 1 COMPLETED

Job ID	Job Name	Status	Stages	Tasks	Submission Time	Duration
3	toPandas	COMPLETED	2/2	388 / 388	4 minutes ago	36s

Visualize with seaborn

In [19]:
```
# heatmap of service availability
plt.figure(figsize=(10, 6))
ax = sns.heatmap(df_loadAvg_pandas.pivot(index='submitter_host', columns='hr', values='avg'), cmap="Blues")
ax.set_title("Heatmap of loadAvg")
```

Out[19]: Text(0.5,1,u'Heatmap of loadAvg')



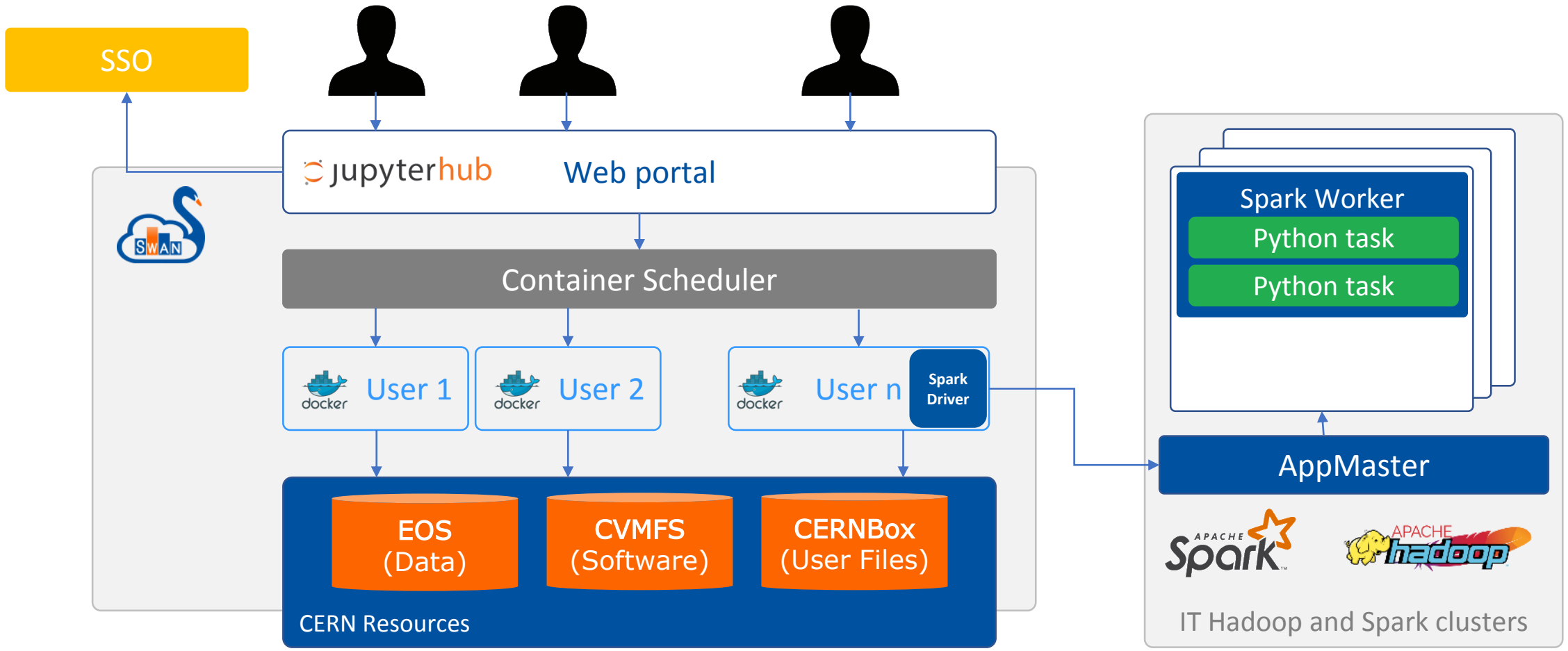
Text

Code

Monitoring

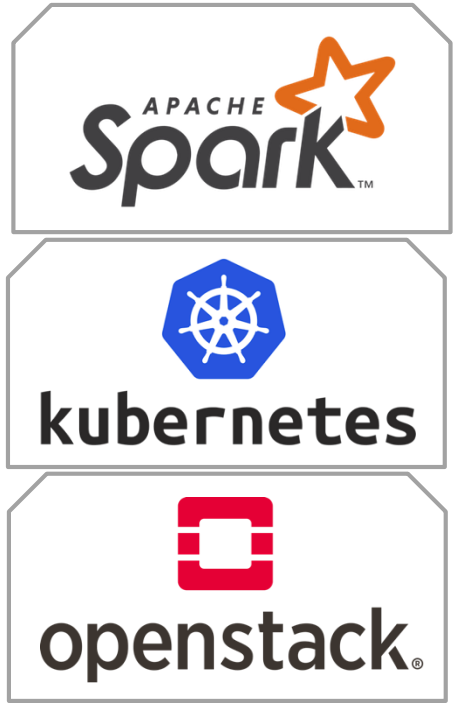
Visualizations

SWAN_Spark – Architecture



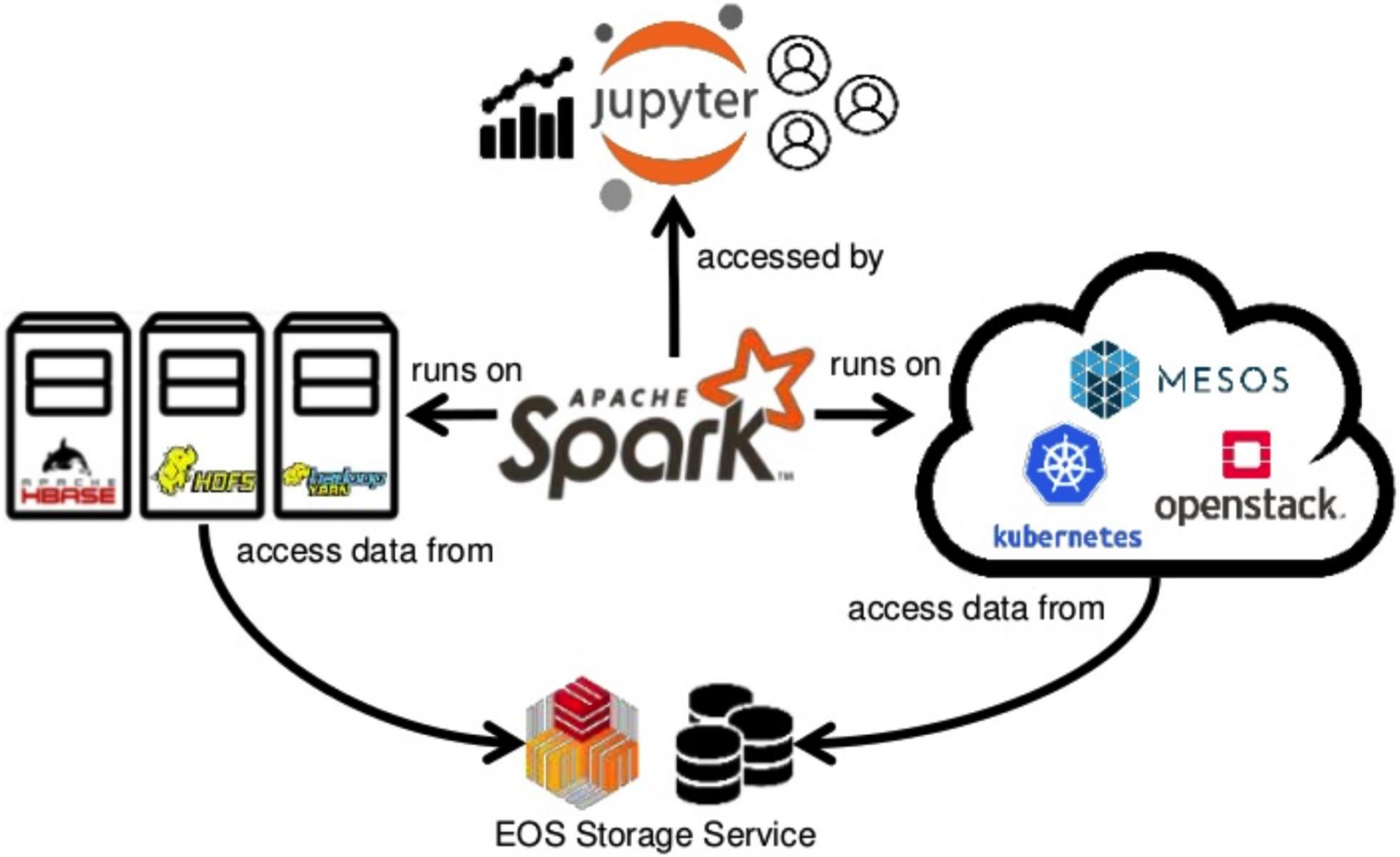
Spark as a service on a private cloud

- Under R&D since 2018
- Appears to be a good solution when data locality is not needed
  - CPU and memory intensive rather than IO intensive workloads
  - Reading from external storages (AFS, EOS, foreign HDFS)
  - Compute resources can be flexibly scale out
- Spark clusters - on containers
  - Kubernetes over Openstack
  - Leveraging the Kubernetes support in Spark 2.3
- Use cases
  - Ad-hoc users with high demand computing resource demanding workloads
  - Streaming jobs (e.g. accessing Apache Kafka)



Analytics platform outlook

- High throughput IO and compute workloads
- Established systems

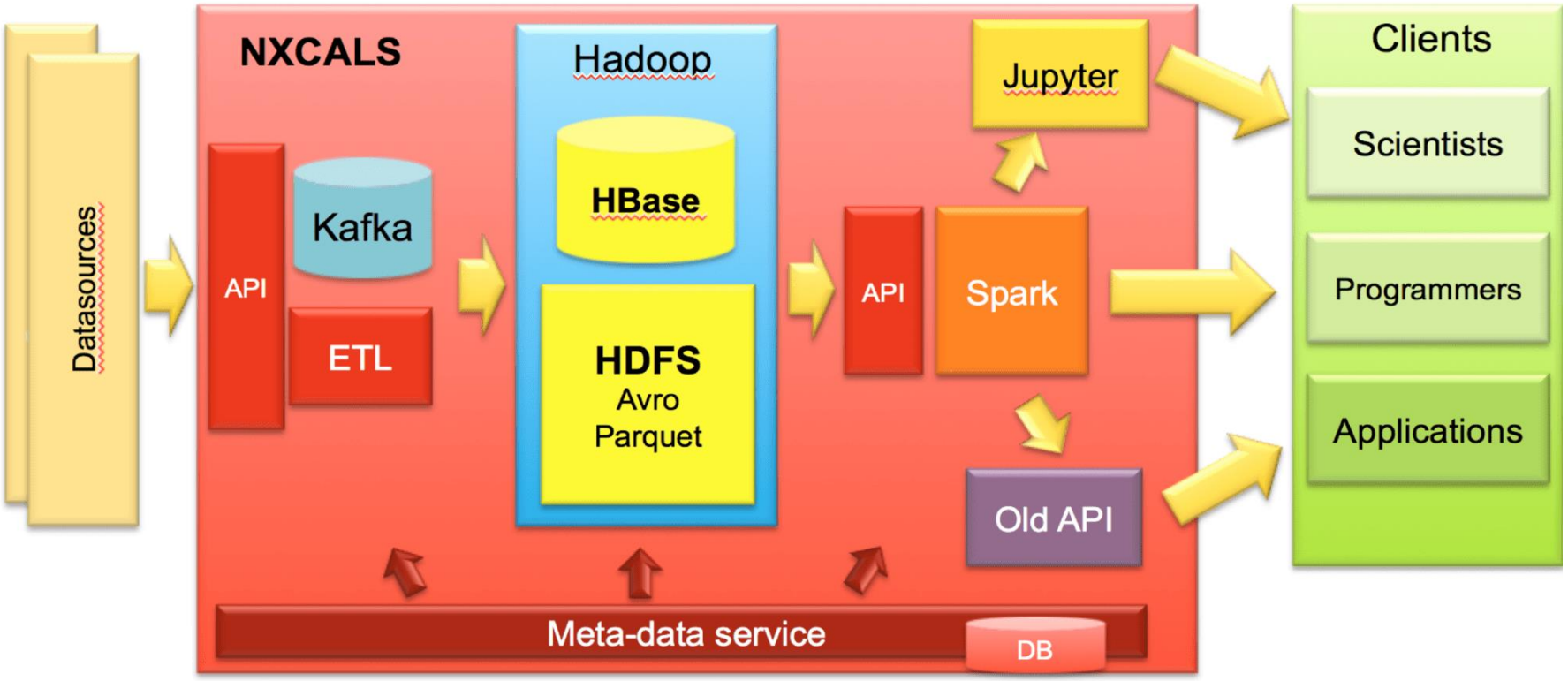


- Flexible scalability for compute intensive workloads
- Ad-hoc users

Selected “Big Data” Projects at CERN

Next Gen. CERN Accelerator Logging

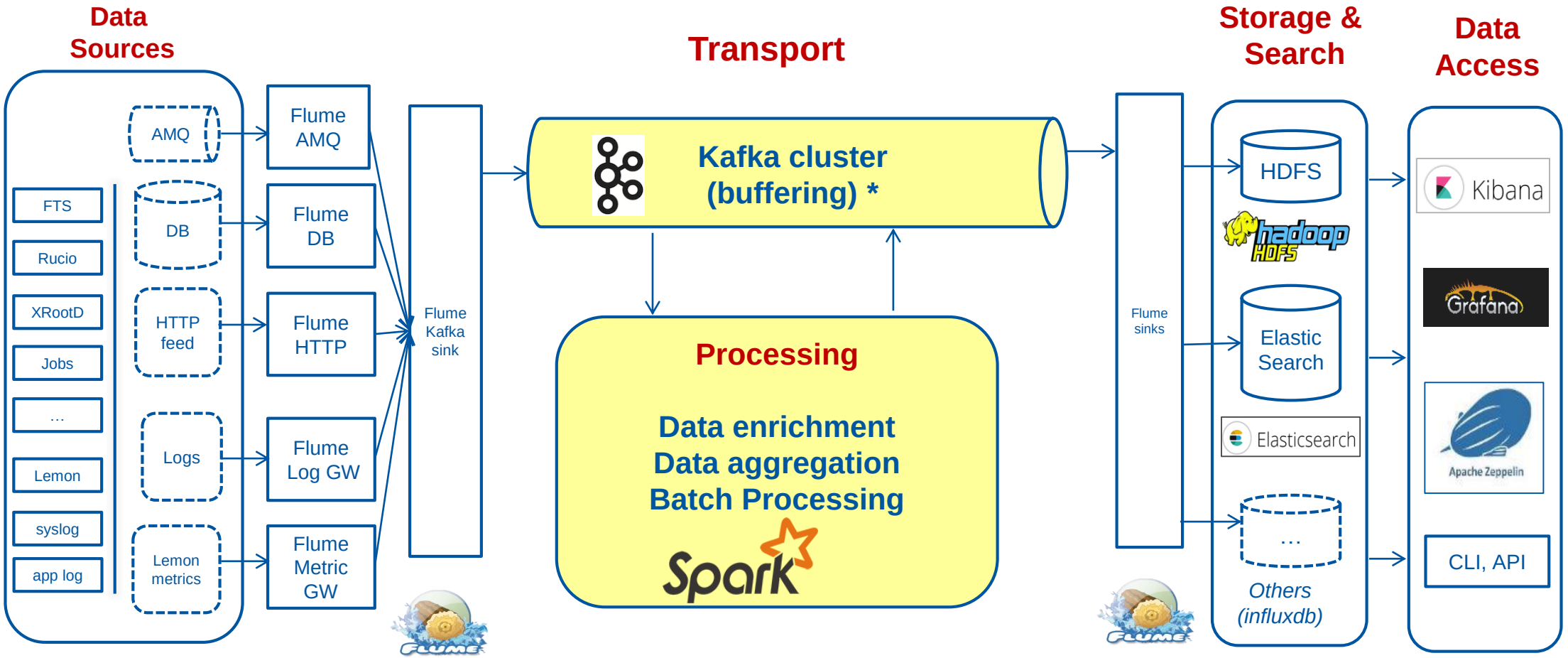
- A control system with: Streaming, Online System, API for Data Extraction
- Critical system for running LHC - 700 TB today, growing 200 TB/year
- Challenge: service level for critical production



Credit: Jakub Wozniak, BE-CO-DS

New CERN IT Monitoring infrastructure

Critical for CC operations and WLCG

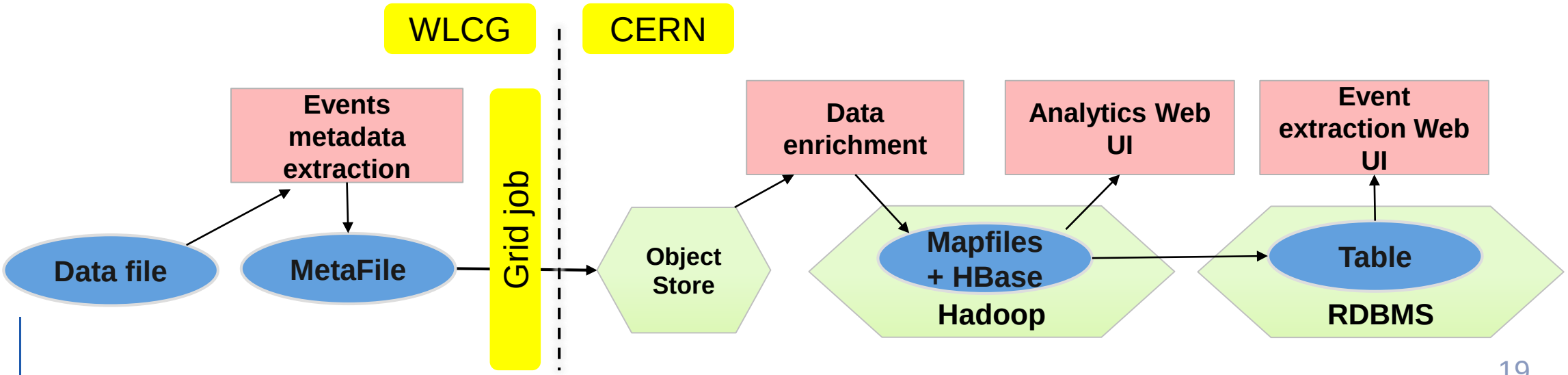


- Data now 200 GB/day, 200M events/day
- At scale 500 GB/day
- Proved to be effective in several occasions

Credits: Alberto Aimar, IT-CM-MM

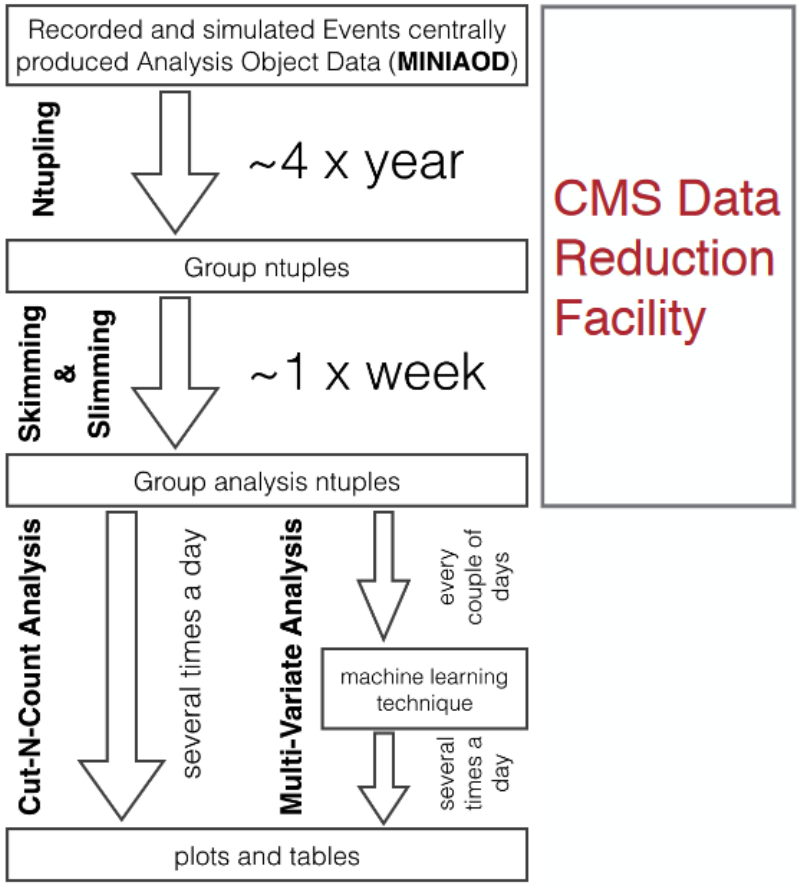
The ATLAS EventIndex

- Catalogue of all collisions in the ATLAS detector
  - Over 120 billion of records, 150TB of data
  - Current ingestion rates 5kHz, 60TB/year



CMS Data Reduction Facility

- R&D, CMS Bigdata project, CERN openlab, Intel:
  - Reduce time to physics for PB-sized datasets
  - Exploring a possible new way to do HEP analysis Improve computing resource utilization
  - Enable physicists to use tools and methods from "Big Data" and open source communities
- CMS Data Reduction Facility:
  - Goal: produce reduced data n-tuples for analysis in a more agile way than current methods
  - Currently testing: scale up with larger data sets, first prototype successful but only used 1TB



R&D: Data Analysis with PySpark for HEP

ROOT
Data Analysis Framework

EOS Storage Service

WLCG
Worldwide LHC Computing Grid

- Data access – **XRootD**
- Reading root files – **spark-root**
- User interface - **SWAN**

APACHE Spark™

hadoop

jupyter

SWAN

Data Ingestion: spark-root

0.1.16 available on Maven Central!

- **spark-root** - ROOT I/O for JVM
- Extends Apache Spark's Data Source API
- Maps each ROOT TTree to Dataset[Row]
- Parallelization = # ROOT files.

Credits: V. Khristenko, J. Pivarski, diana-hep and CMS Big Data project

Scala

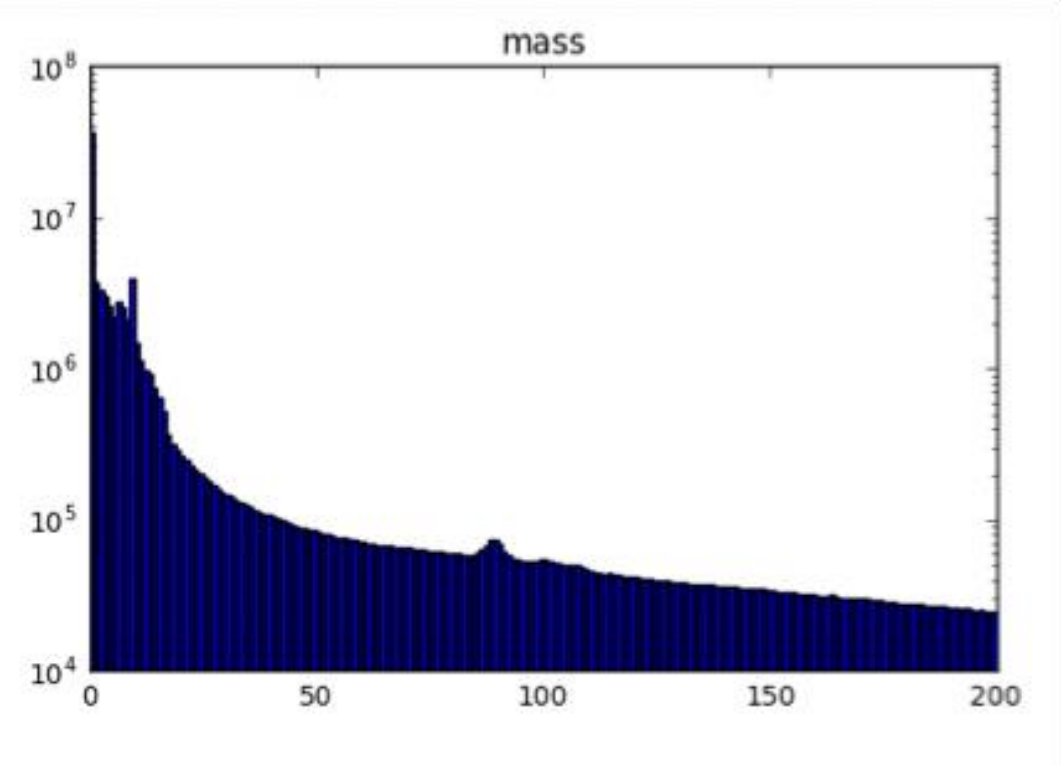
```
// inject the Dataset[Row]
import org.dianahep.sparkroot.experimental._
val df = spark.read.option("tree", <treeName>).root("<path/to/file>")

// pretty print of the schema
df.printSchema
```

```
|-- Particle: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- fUniqueID: integer (nullable = true)
 |    |    |-- fBits: integer (nullable = true)
 |    |    |-- PID: integer (nullable = true)
 |    |    |-- Status: integer (nullable = true)
 |    |    |-- IsPU: integer (nullable = true)
 |    |    |-- M1: integer (nullable = true)
 |    |    |-- M2: integer (nullable = true)
 |    |    |-- D1: integer (nullable = true)
 |    |    |-- D2: integer (nullable = true)
 |    |    |-- Charge: integer (nullable = true)
 |    |    |-- Mass: float (nullable = true)
 |    |    |-- E: float (nullable = true)
```

Data Processing: CMS Open Data Example

Let's calculate the invariant mass of a di-muon system?!

- Transform a collection of muons to an invariant mass for each Row (Event).
- Aggregate (histogram) over the entire dataset.




```
# read in the data
df = sqlContext.read\
                    .format("org.dianahep.sparkroot.experimental")\
                    .load("hdfs:/path/to/files/*.root")

# count the number of rows:
df.count()

# select only muons
muons =
df.select("patMuons_slimmedMuons__RECO_.patMuons_slimmedMuo
ns__RECO_obj.m_state").toDF("muons")

# map each event to an invariant mass
inv_masses = muons.rdd.map(toInvMass)

# Use histogrammar to perform aggregations
empty = histogrammar.Bin(200, 0, 200, lambda row: row.mass)
h_inv_masses = inv_masses.aggregate(empty,
                    histogrammar.increment,
                    histogrammar.combine)
```

Credits: V. Khristenko, J. Pivarski, diana-hep and CMS Big Data project

Conclusions

- Demand of “Big Data” platforms and tools is growing at CERN
  - Many projects started and running
  - Projects around Monitoring, Security, Accelerators logging/controls, physics data, streaming…
- Hadoop, Spark, Kafka services at CERN IT
  - Service is evolving: High availability, security, backups, external data sources, notebooks, cloud…
- Experience and community
  - Technologies evolve rapidly and knowledge sharing very important
  - We are happy to share/exchange our experience, tools, software with others
  - HEP sites welcome to join discussions at Hadoop User Forum: https://indico.cern.ch/category/5894/
  - Interest of HEP sites in collaborating with CERN on some projects?

Future work

- Spark on Kubernetes (continuation)
- Rolling out Kafka service in full production state
- Improve monitoring of individual users in Hadoop
- Service web portal for users
  - Integrate multiple components of the service
  - Single place for monitoring and requesting the service resources
    - HDFS quota, CPUs, memory, Kafka topics etc
- Explore further the big data technology landscape
  - Presto, Apache Kudu, Apache Beam etc.

Acknowledgements

- Colleagues in the Hadoop, Spark and Streaming Service at CERN
- Users of the service, including IT Monitoring, IT Security, ATLAS DDM and EventIndex, CMSSpark project, Accelerator logging and controls
- CMS Big Data project, with Intel and openlab
- Collaboration on SWAN with CERN colleagues in EP-SFT, IT-ST