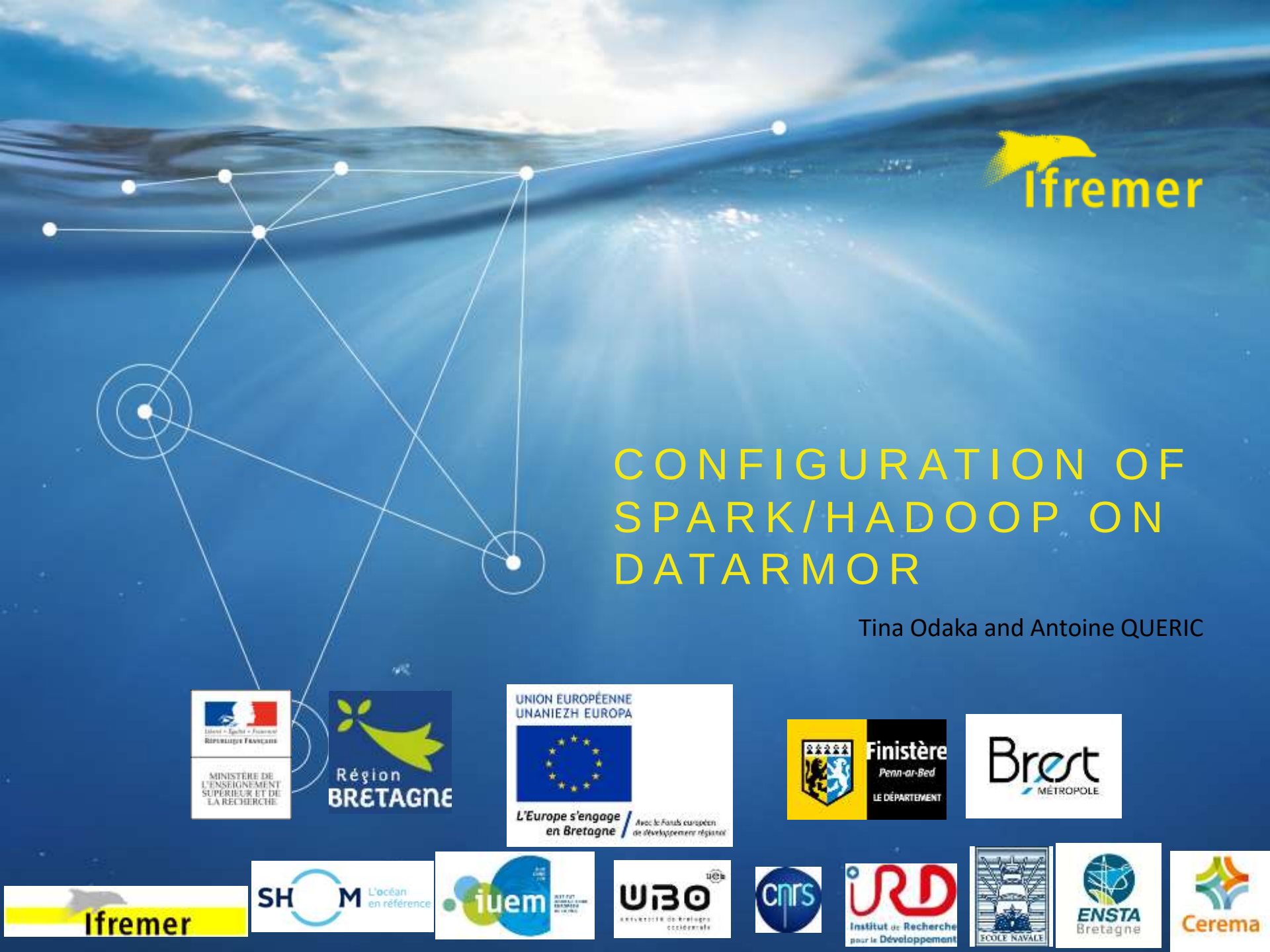
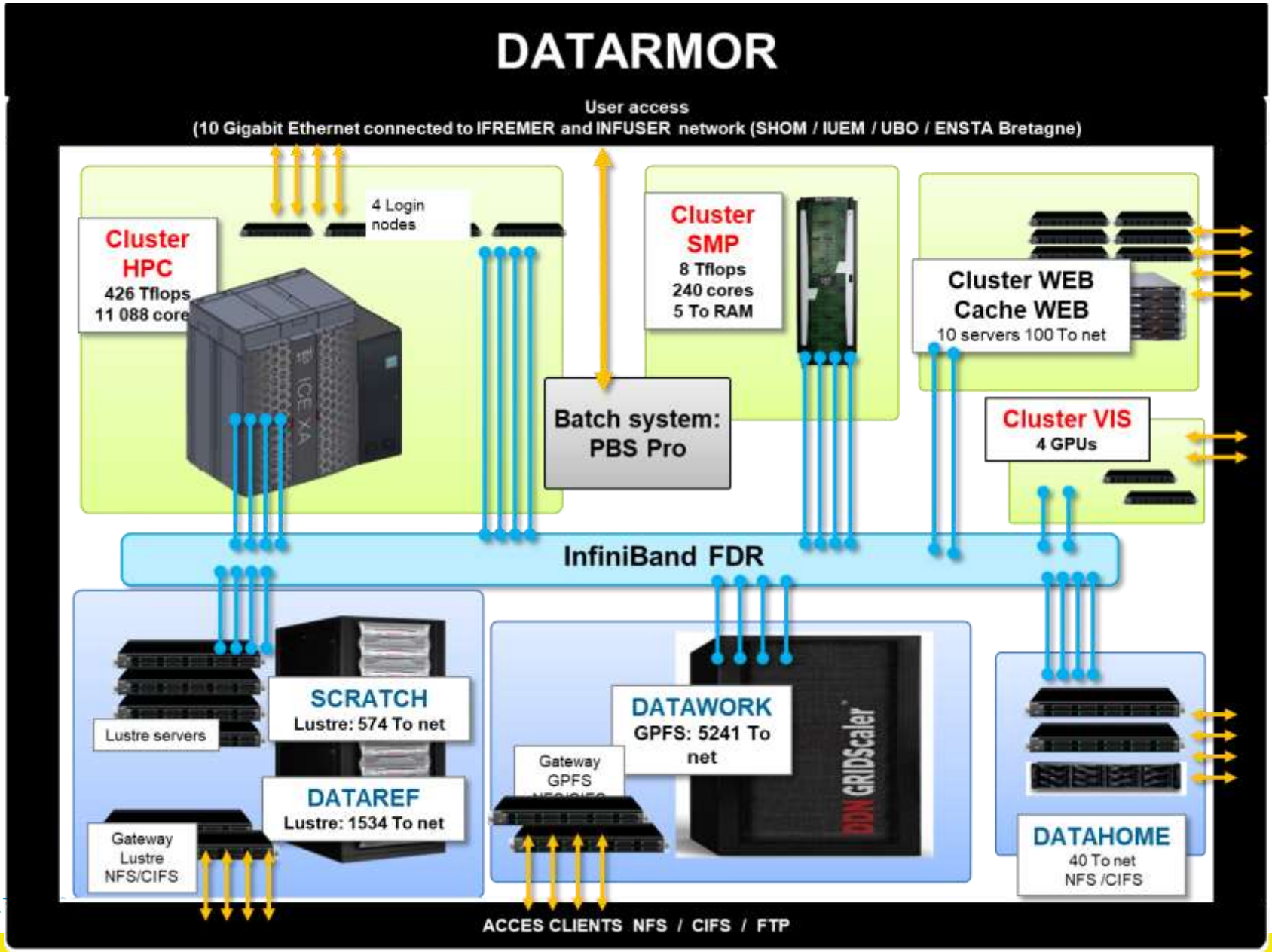


CONFIGURATION OF SPARK/HADOOP ON DATARMOR

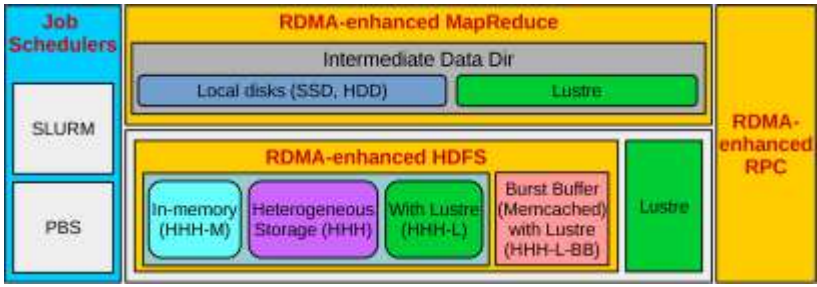
Tina Odaka and Antoine QUERIC


DATARMOR
User access
(10 Gigabit Ethernet connected to IFREMER and INFUSER network (SHOM / IUEM / UBO / ENSTA Bretagne)
4 Login nodes
Cluster HPC
426 Tflops
11 088 core
Cluster SMP
8 Tflops
240 cores
5 To RAM
Cluster WEB
Cache WEB
10 servers 100 To net
Cluster VIS
4 GPUs
Batch system: PBS Pro
InfiniBand FDR
SCRATCH
Lustre: 574 To net
Lustre servers
DATAREF
Lustre: 1534 To net
Gateway Lustre NFS/CIFS
Gateway GPFS
DATAWORK
GPFS: 5241 To net
DDN GRIDScaler
DATAHOME
40 To net
NFS /CIFS
ACCES CLIENTS NFS / CIFS / FTP

DATARMOR is not a ‘standard’ Hadoop cluster

- Network connections is InfiniBand FDR, 56 times faster than 1Gs ‘standard network’, and have RDMA (Remote Direct Memory Access) capability.
- Cluster HPC does not have local disk to form ‘standard HDFS’.
- DATARMOR have two different parallel file system LUSTRE, and GPFS (IBM Spectrum Scale)

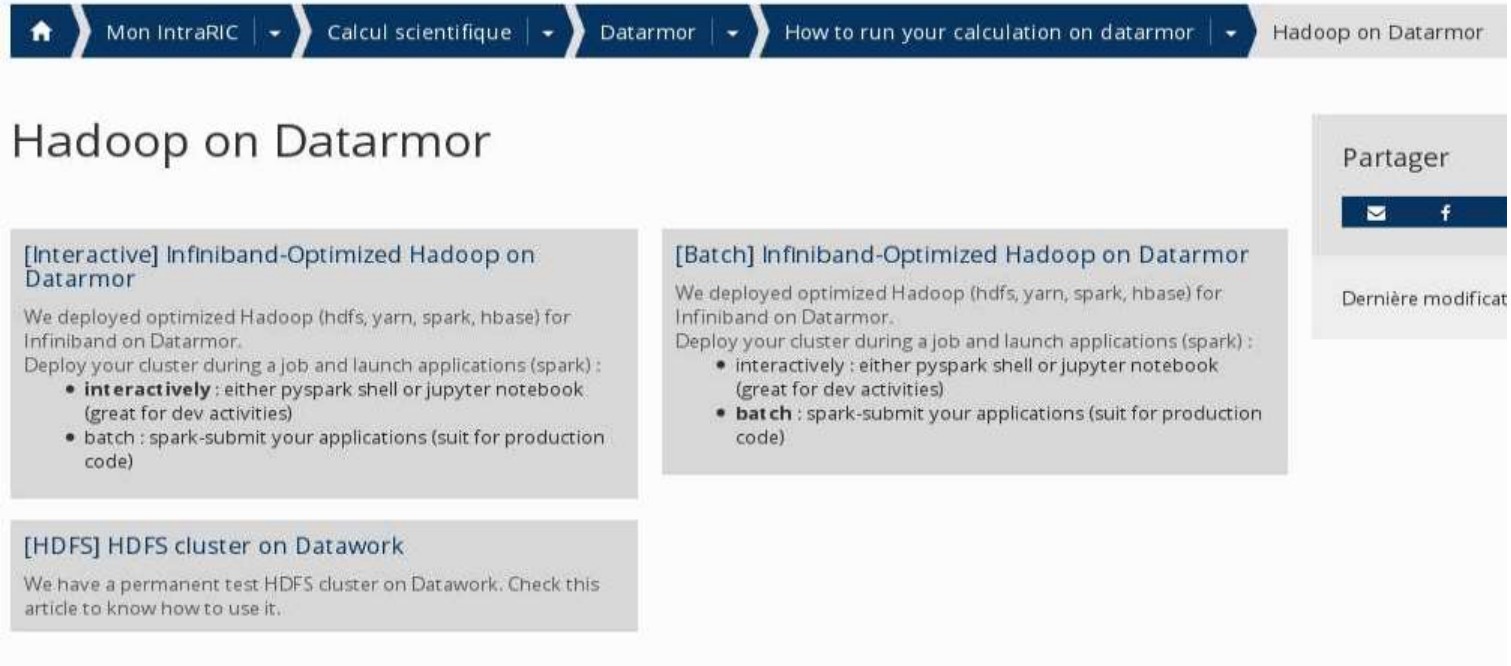
- Infiniband optimized:
  RDMA for Apache Spark, Hadoop, and Hbase High-Performance BigData (HiBD)
  http://hibd.cse.ohio-state.edu



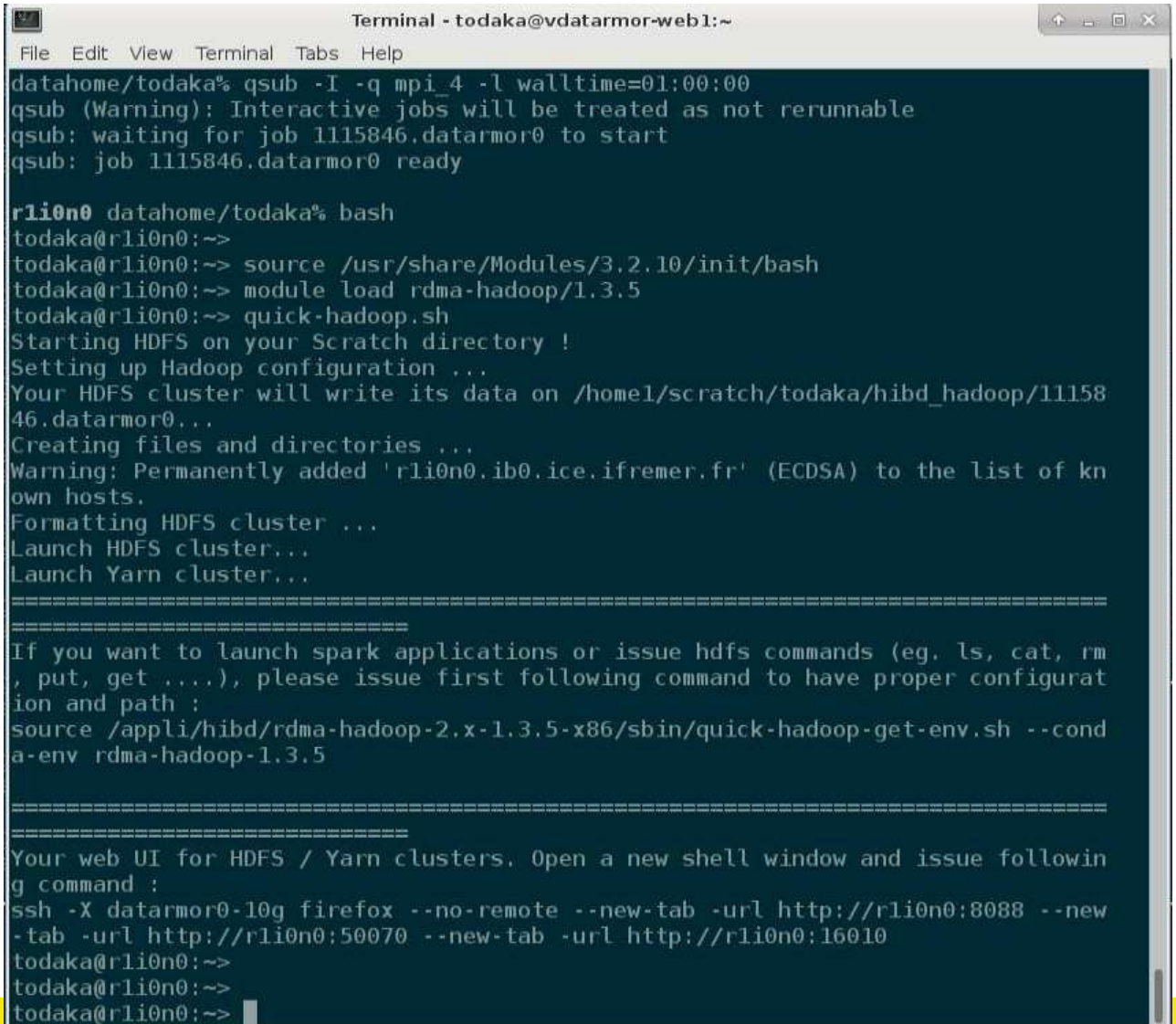
- HDFS transparency
  Spectrum scale, IBM product

In practice;

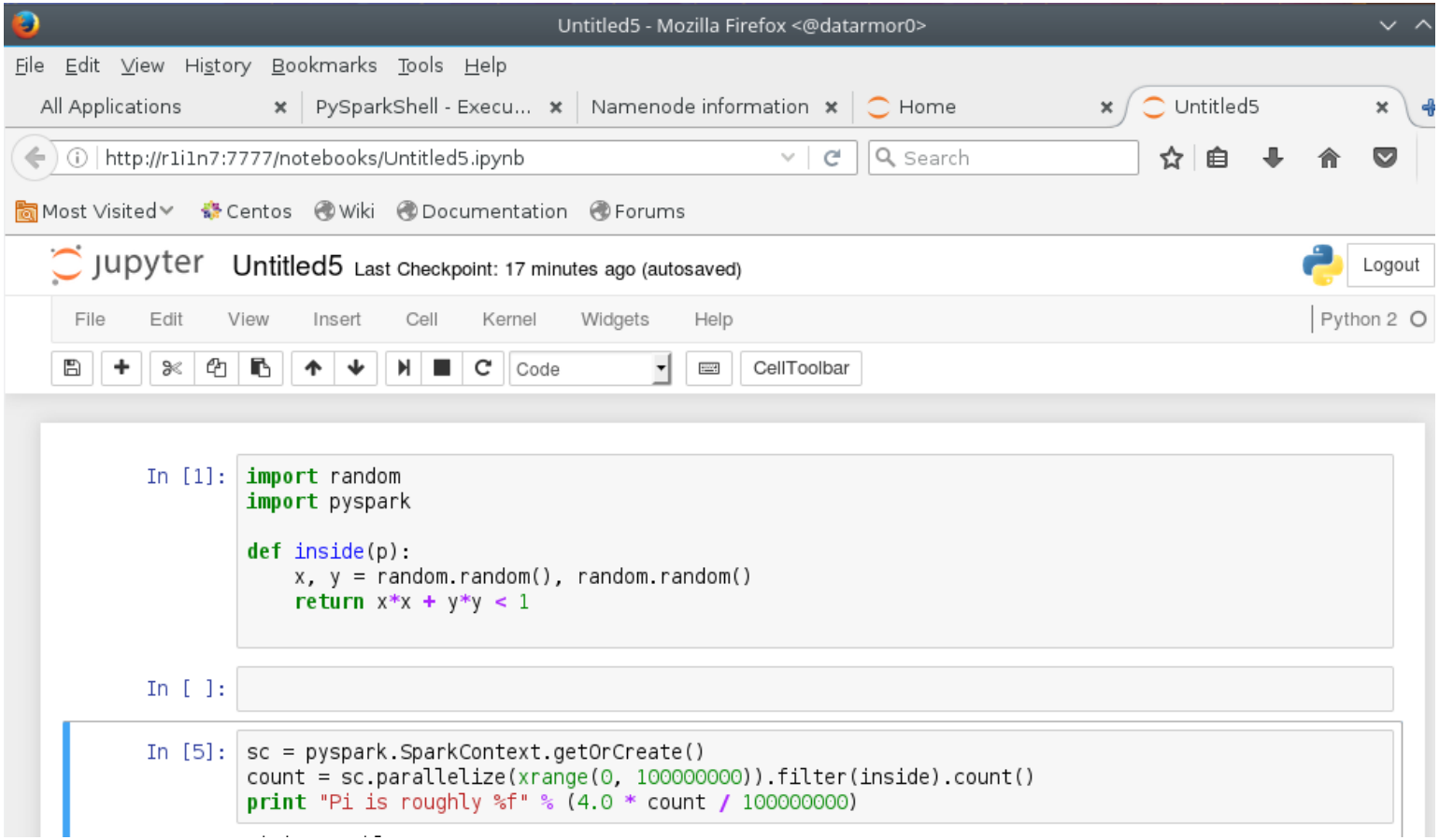
Documentation-> Datarmor –How to run your calculation on datarmor -> Hadoop on Datarmor



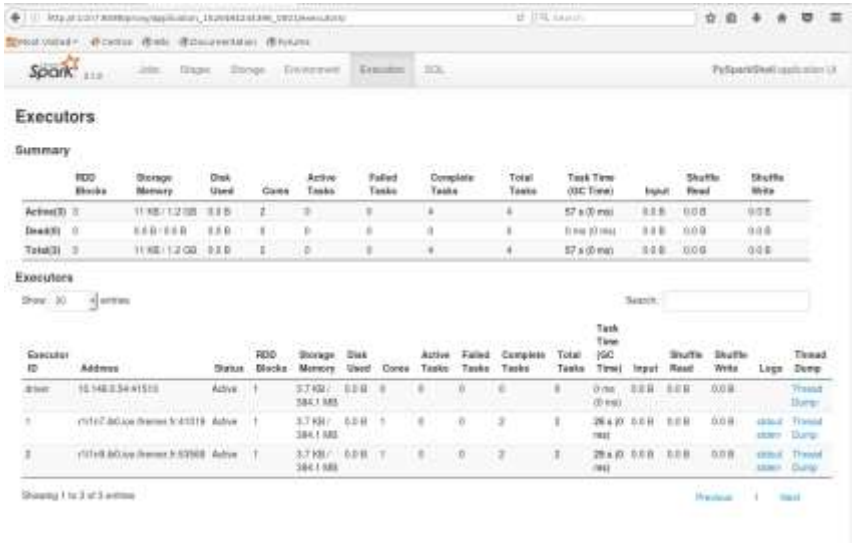
In practice

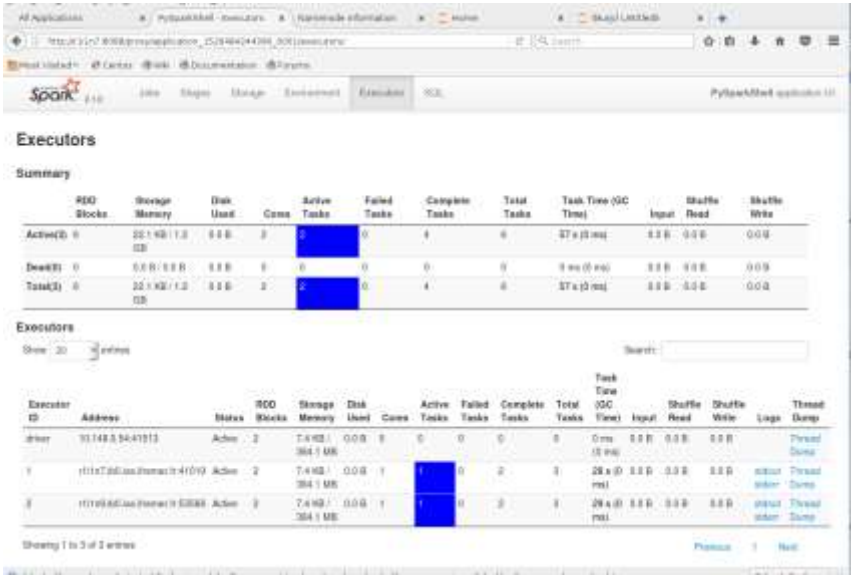

Terminal - todaka@vdatarmor-web1:~
File Edit View Terminal Tabs Help
datahome/todaka% qsub -I -q mpi_4 -l walltime=01:00:00
qsub (Warning): Interactive jobs will be treated as not rerunnable
qsub: waiting for job 1115846.datarmor0 to start
qsub: job 1115846.datarmor0 ready
r1i0n0 datahome/todaka% bash
todaka@r1i0n0:~>
todaka@r1i0n0:~> source /usr/share/Modules/3.2.10/init/bash
todaka@r1i0n0:~> module load rdma-hadoop/1.3.5
todaka@r1i0n0:~> quick-hadoop.sh
Starting HDFS on your Scratch directory !
Setting up Hadoop configuration ...
Your HDFS cluster will write its data on /home1/scratch/todaka/hibd_hadoop/1115846.datarmor0...
Creating files and directories ...
Warning: Permanently added 'r1i0n0.ib0.ice.ifremer.fr' (ECDSA) to the list of known hosts.
Formatting HDFS cluster ...
Launch HDFS cluster...
Launch Yarn cluster...
If you want to launch spark applications or issue hdfs commands (eg. ls, cat, rm, put, get ....), please issue first following command to have proper configuration and path :
source /appli/hibd/rdma-hadoop-2.x-1.3.5-x86/sbin/quick-hadoop-get-env.sh --conda-env rdma-hadoop-1.3.5
Your web UI for HDFS / Yarn clusters. Open a new shell window and issue following command :
ssh -X datarmor0-10g firefox --no-remote --new-tab -url http://r1i0n0:8088 --new-tab -url http://r1i0n0:50070 --new-tab -url http://r1i0n0:16010
todaka@r1i0n0:~>
todaka@r1i0n0:~>
todaka@r1i0n0:~>

Jupyter notebook: pyspark



spark@datarmor





status

Now:

Checking the installation with beta, experienced users;

Next step:

benchmark to verify the performance.