DataScale
Seismic correlation on HPC

Ter@tec 2014

Outline

- DataScale project
- Seismic use-case
- DataScale architecture

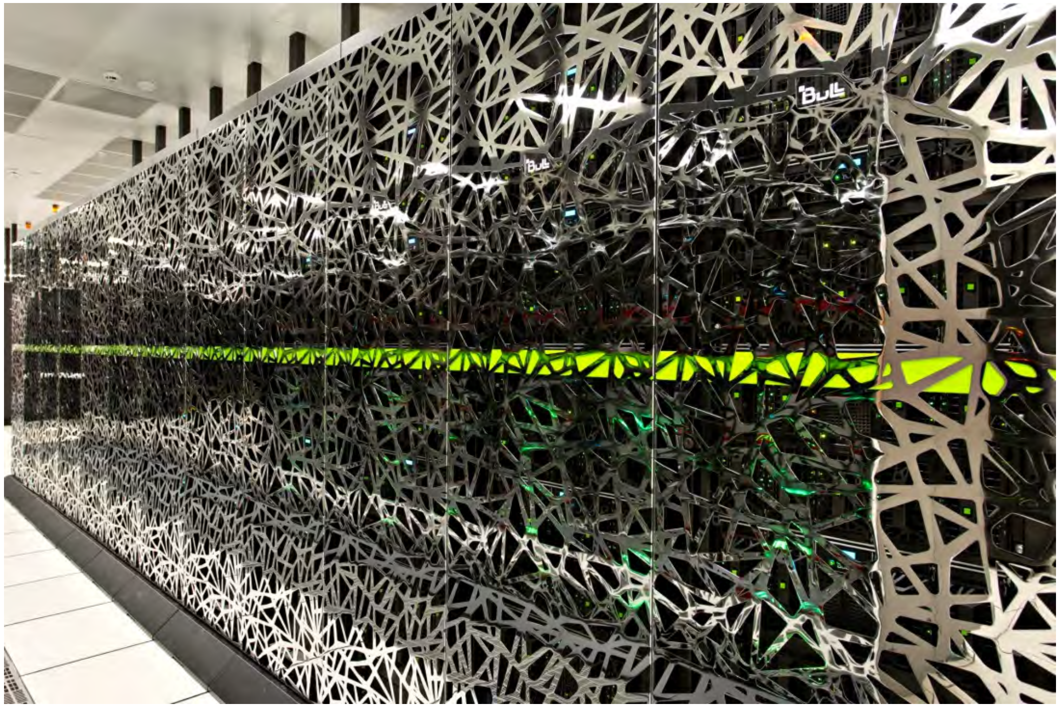
DataScale Project

DataScale Project

- Develop synergies between Big Data and HPC
- Consortium of 9 partners
- Two-year project, started in June 2013
- Supported by the French government
- Funded by the French “Investissements d’Avenir” program

DataScale Project

- Three use-cases
  1. Cluster management
     - Data mining applied to execution logs

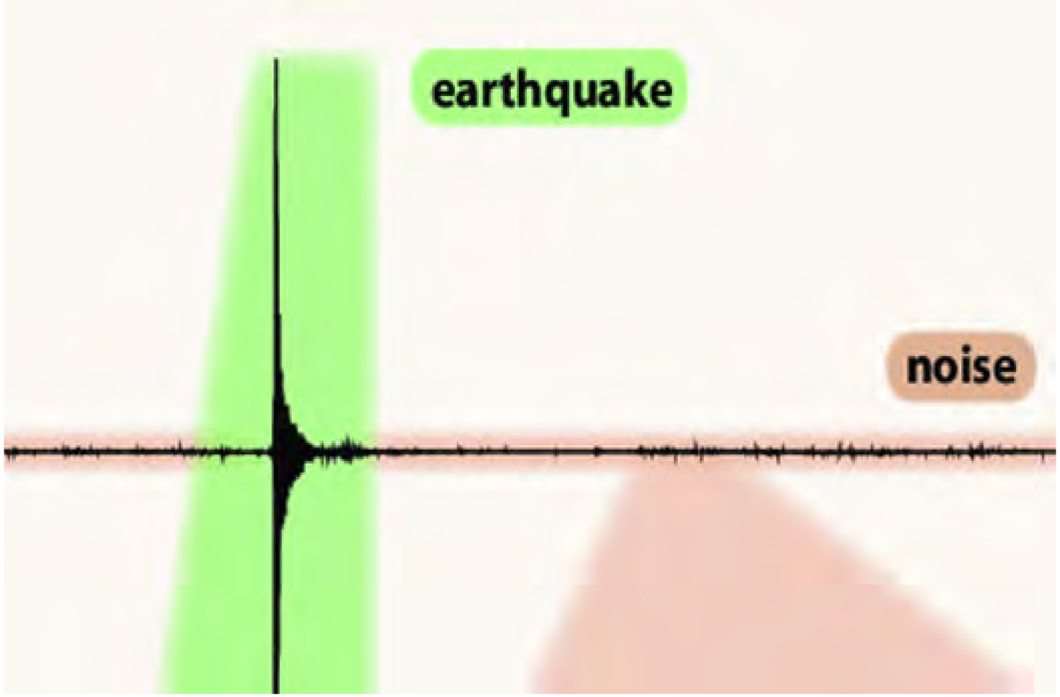


DataScale Project

- Three use-cases
  1. Cluster management
  2. Multimedia product analysis
     - Data mining applied to image search

DataScale Project

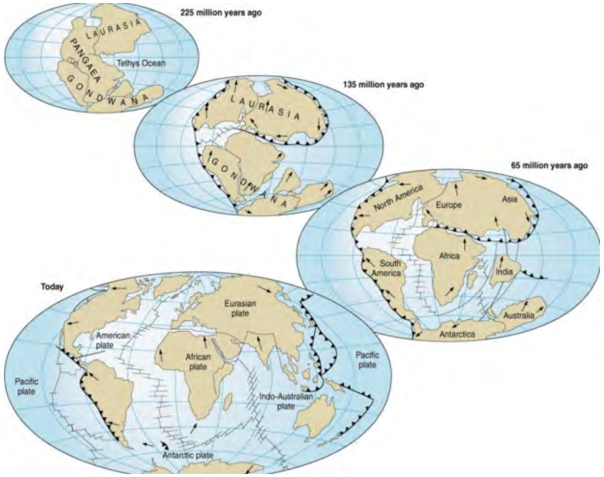
- Three use-cases
  1. Cluster management
  2. Multimedia product analysis
  3. Seismic event detection



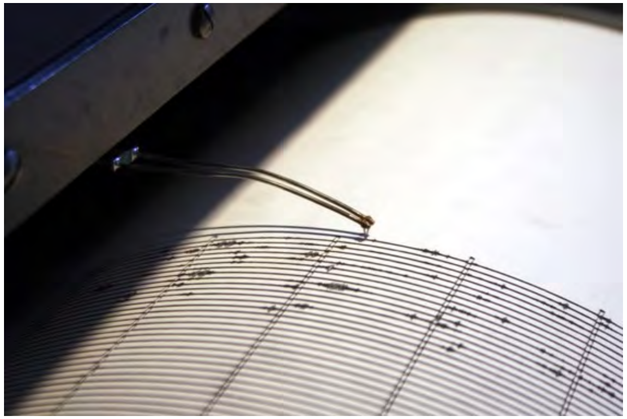
Seismic Use Case

Global seismology

- The Earth is a dynamic system…



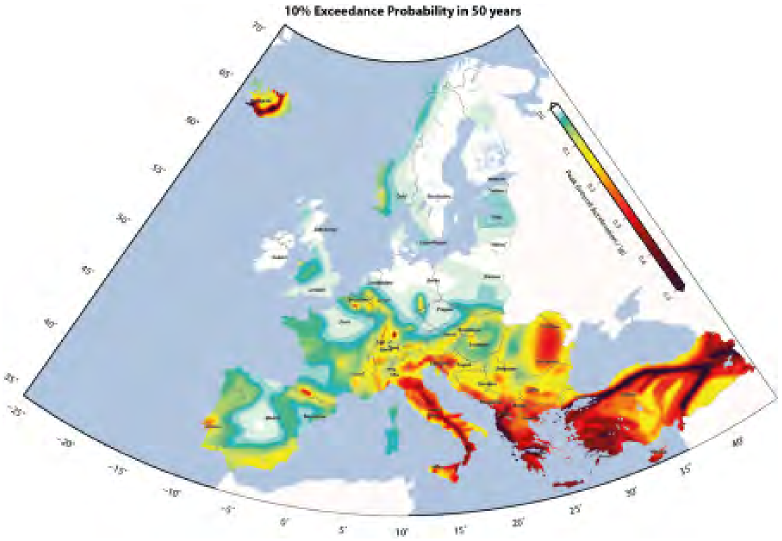
… its pulse is taken continuously



CEA missions

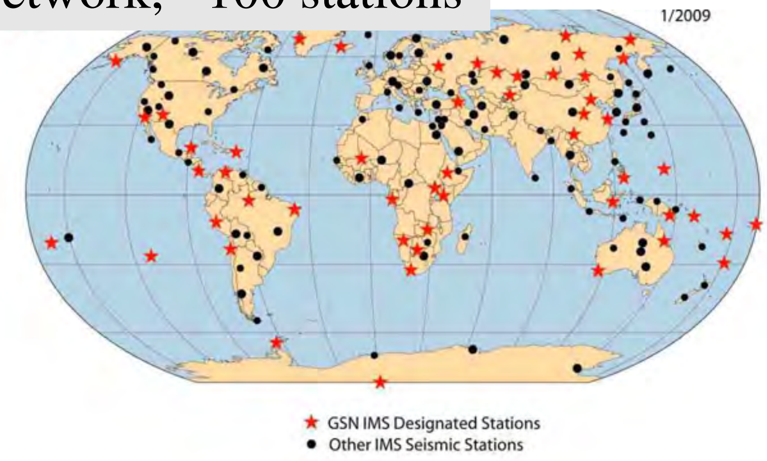
- **Monitoring** and analysis of the seismic activity and **alert** authorities in case of :
  - strong earthquake
  - tsunami
  - nuclear explosion

- Better **understanding** of :
  - phenomenology
  - seismic hazard and seismic risk
  - geologic structure

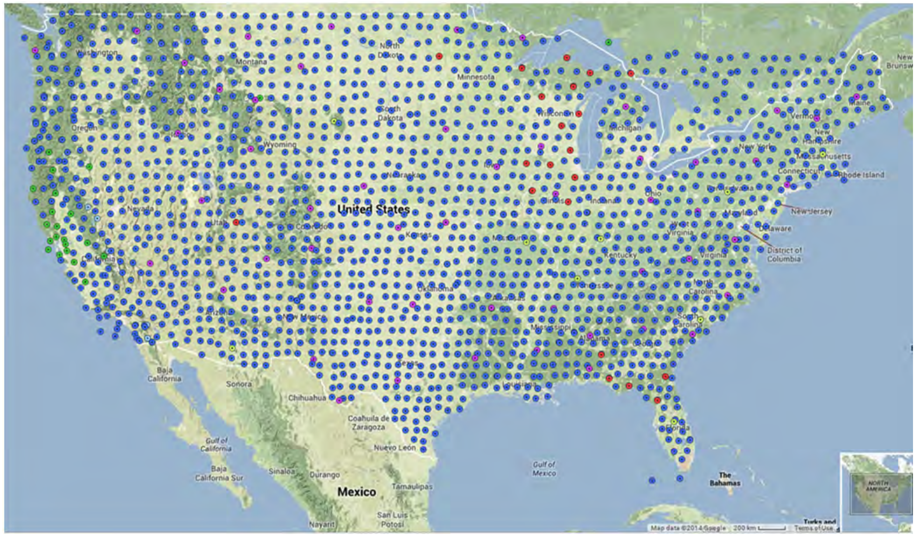


Global and dense networks

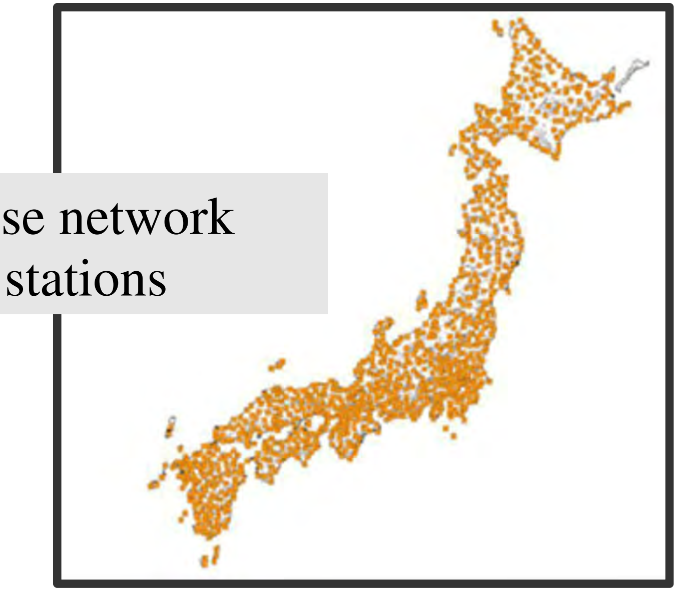
IMS Network, ~100 stations



US Array, 4000 stations

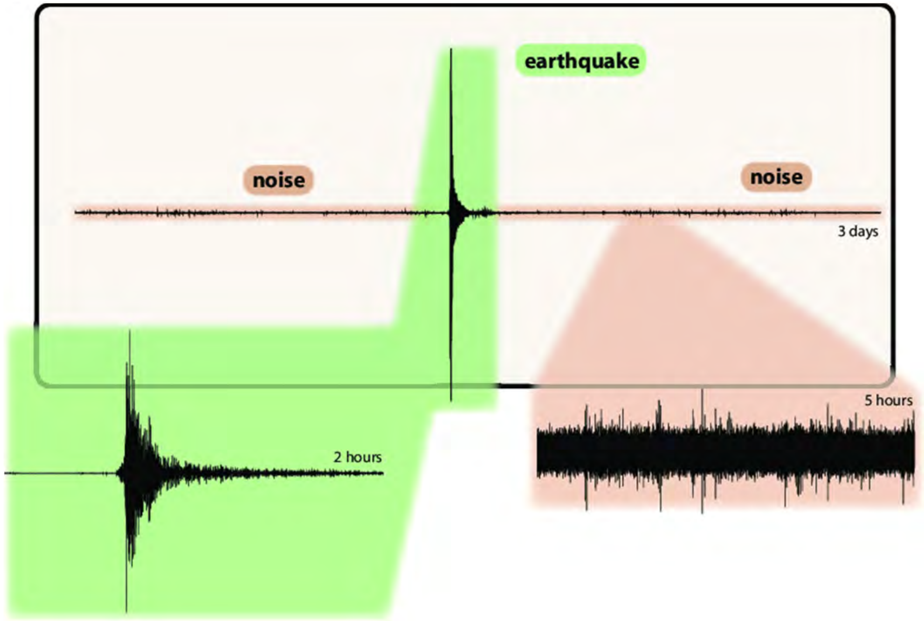


Japanese network
~1000 stations



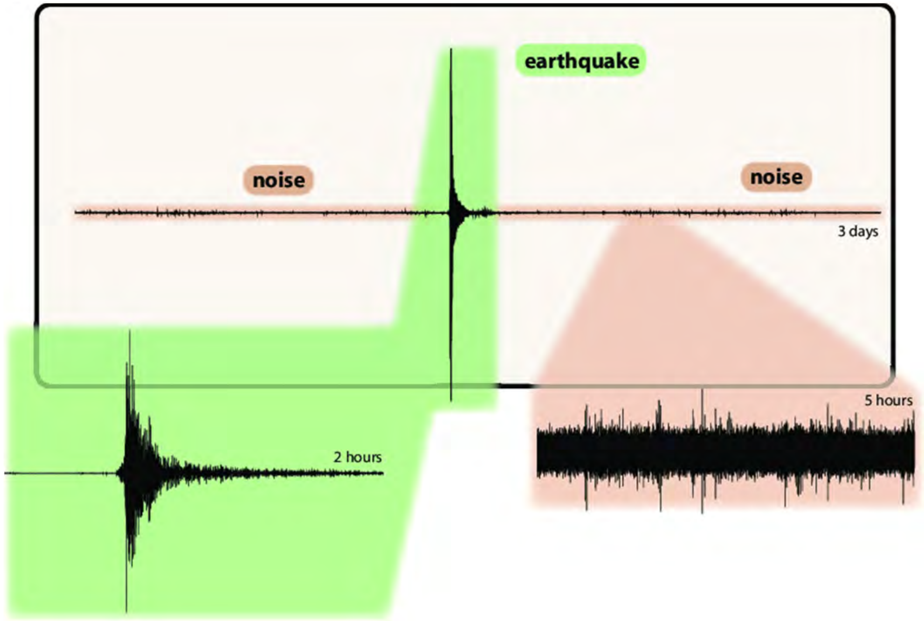
Seismic record

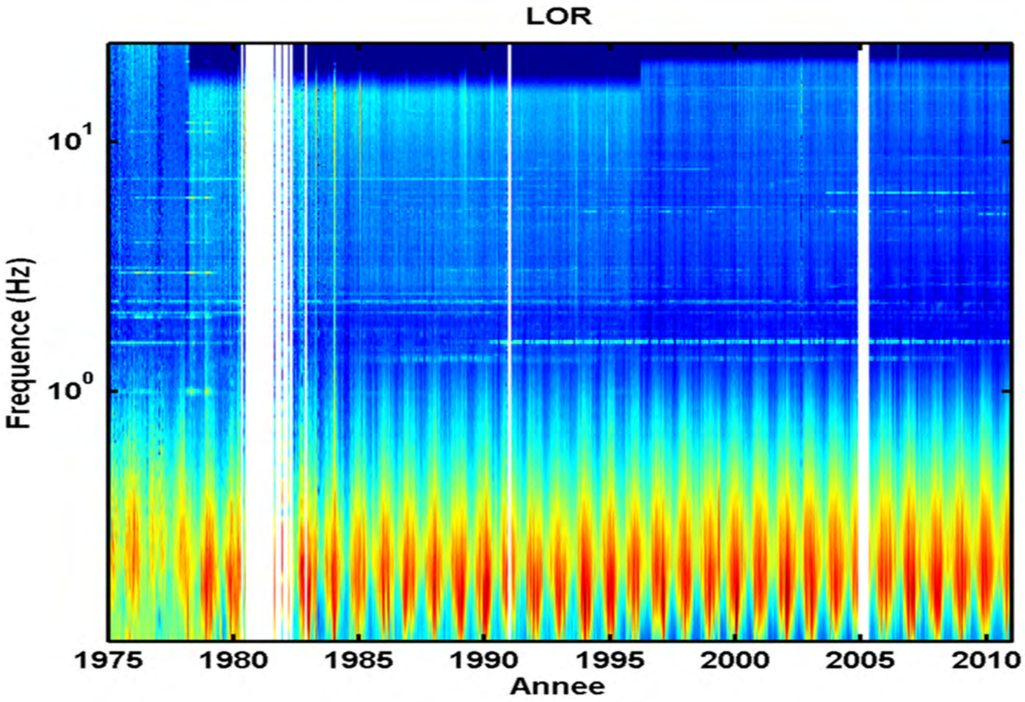
- Data centers collect, process, analyze, produce data 24 hours a day, 7 days a week



Seismic record

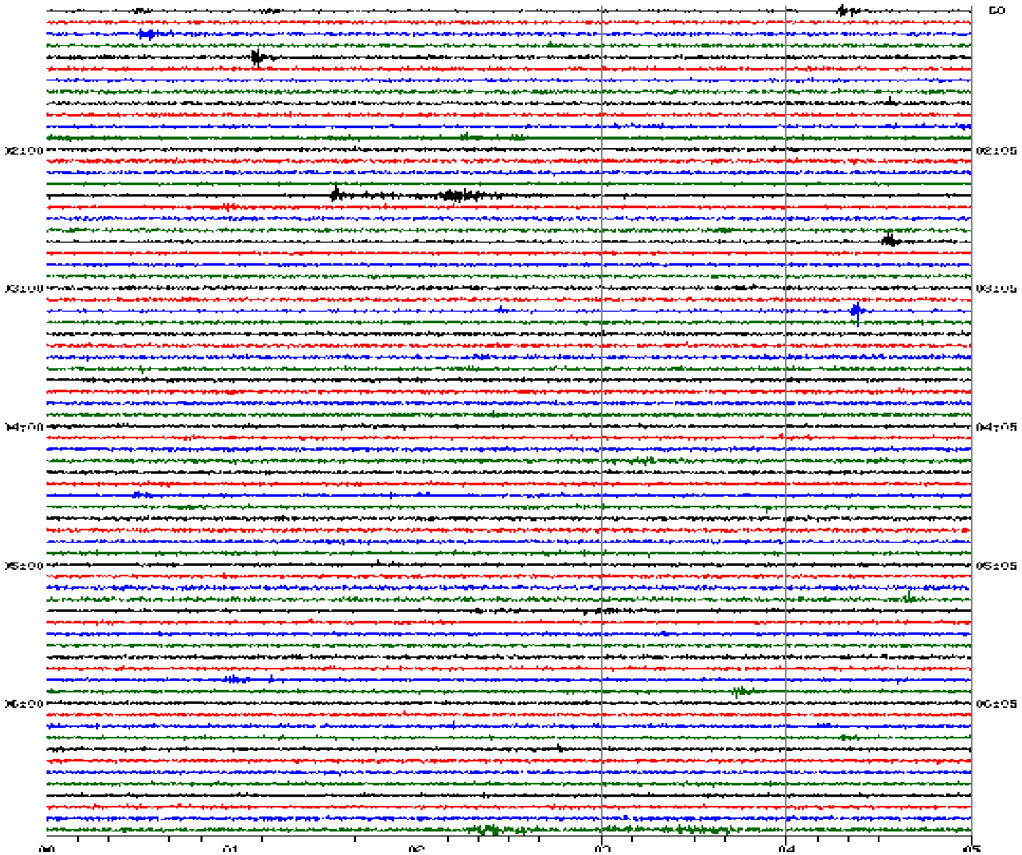
- Data centers collect, process, analyze, produce data 24 hours a day, 7 days a week
- Data is the cornerstone : full of **information** and source of **knowledge**





Millions of seismic records

- Data centers collect, process, analyze, produce data 24 hours a day, 7 days a week
- Data is the cornerstone : full of **information** and source of **knowledge**

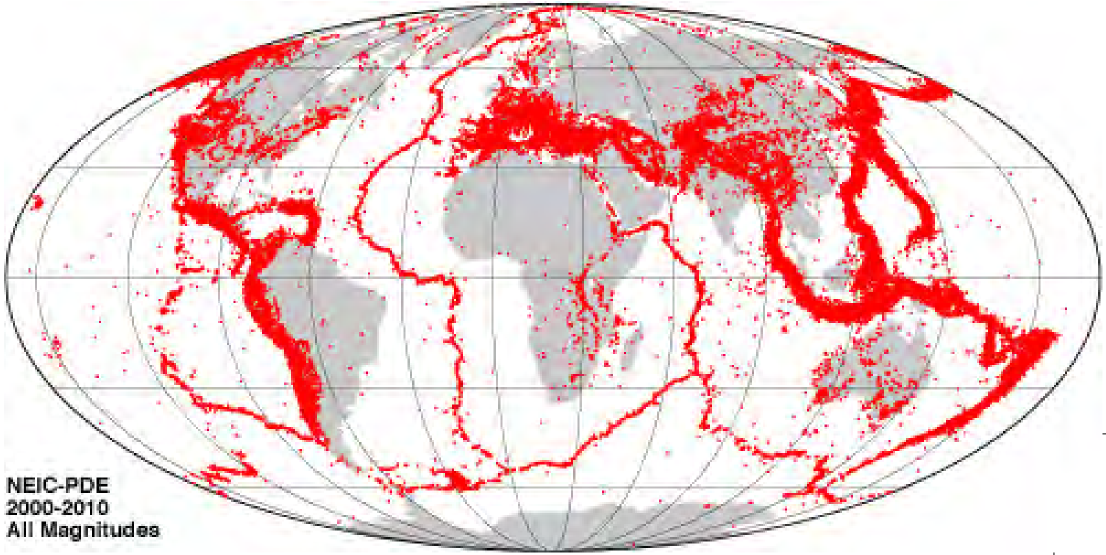


© DataScale

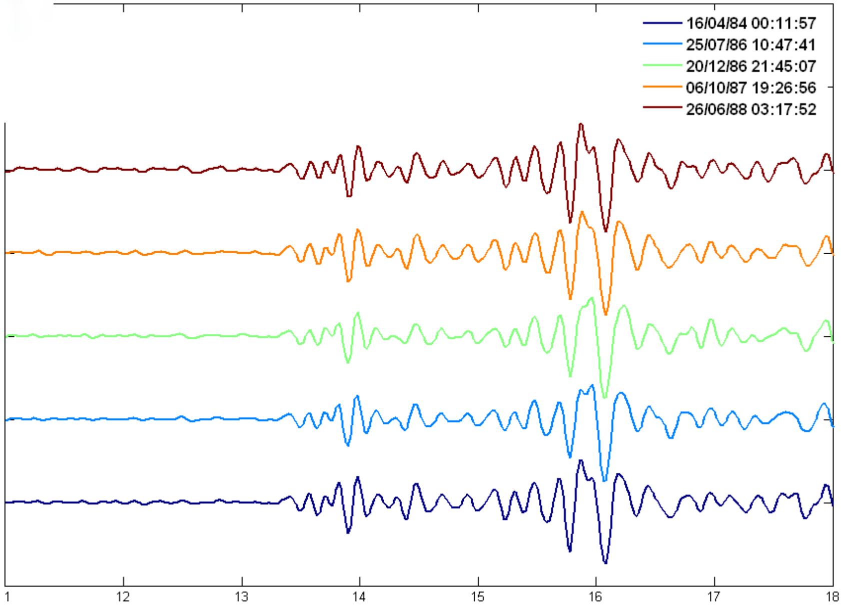
Use case : earthquake detection

ABUNDANT SEISMIC ACTIVITY …

~50.000 events / year (IDC)

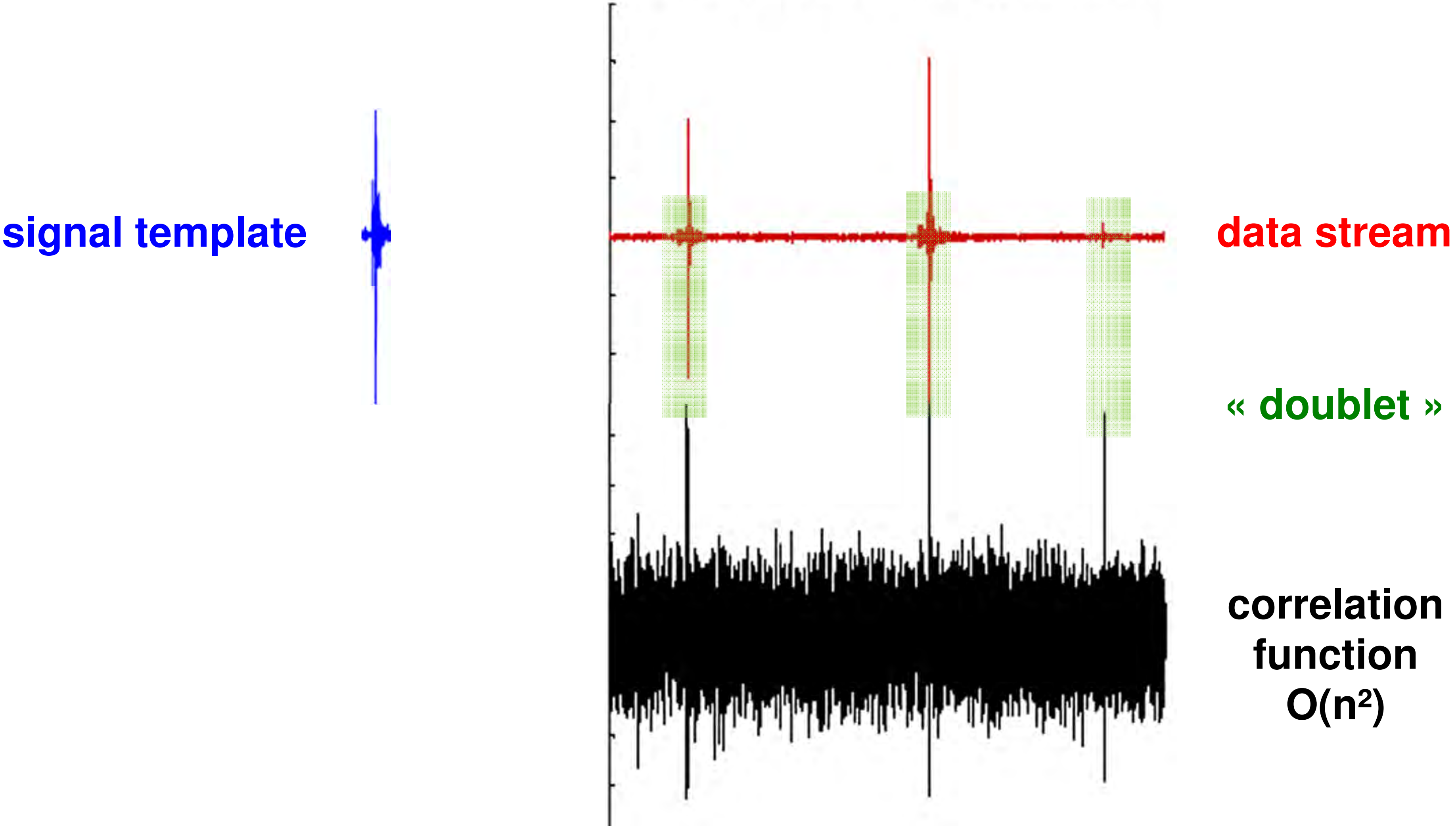


SOMETIMES REDUNDANT

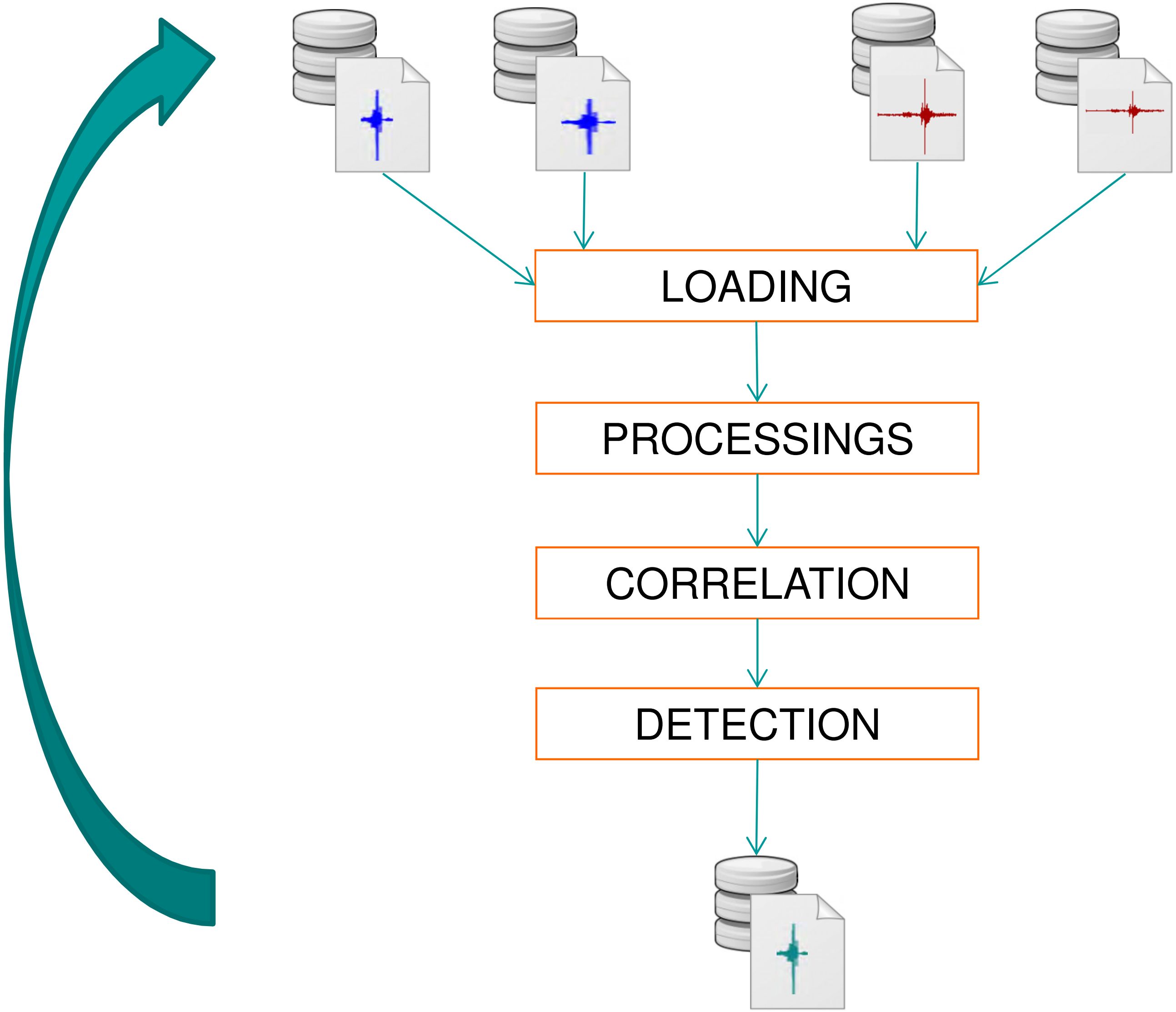


Detection with correlation

- Detection of earthquake using waveform correlation
- Correlation between a template and the (incoming) data stream



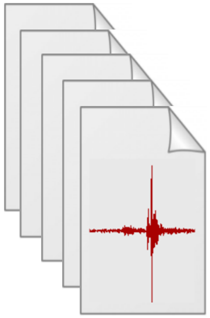
Workflow



© DataScale

Big Data

Data sets are



- **Large and growing**
- **Complex and heterogeneous**
- **Continuous stream and real time**
- **Sometimes imprecise**

Volume
Variety
Velocity
Veracity

I/O 30TB (10 years)
2M files / 500.000 events
100 sensors

A (big) technological problem

- Intrinsic **mismatch between Data and IT** (Information Technology)
- Difficult to process all the data with traditional applications within tolerable elapsed time
- Need of hardware and software solutions

Challenges

- Efficient **data processing**
  - ➔ Distribute, parallelize and deploy the application on HPC platform
- Efficient **data management**
  - ➔ Define hierarchy of data storage (data life cycle, reuse process)
- **New database management** system with data mining technologies
  - ➔ Handle very large volumes and different types of data
- **Extend data and result access**
  - ➔ Open the HPC platform to the cloud

DataScale Architecture

Big Data & DataScale overview

Big Data constraints

- Costly storage for large amount of data
- Need of new methods of data indexation and data mining on large clusters
- Cloud usage is generalized

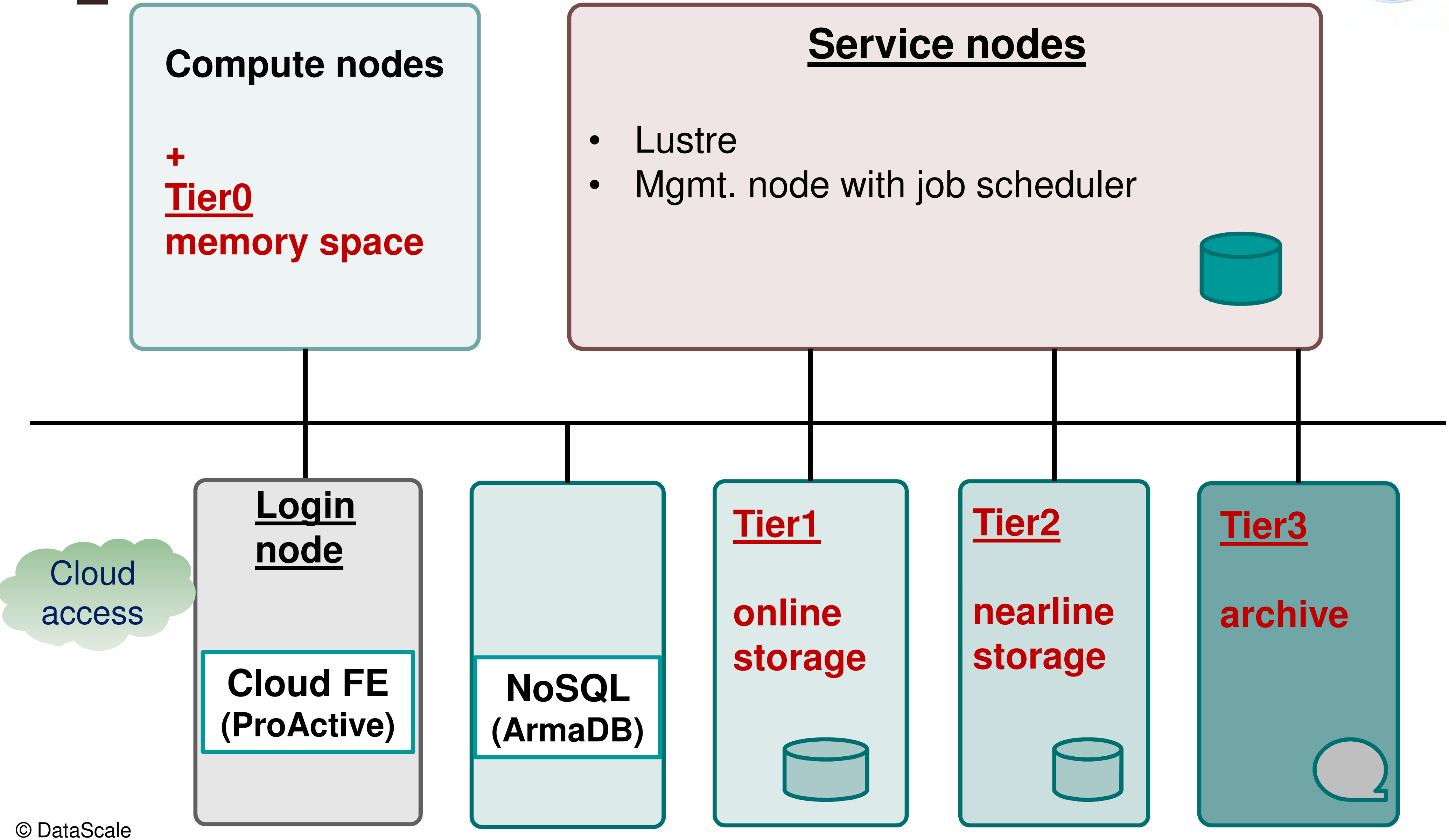
DataScale answers

- Efficient data management
  - ○ Hierarchical storage (HSM)
  - ○ Manage data movement
- NoSQL DB and data mining on HPC nodes
- HPC has to be open to cloud

Software stack

- ProActive
  - Cloud storage and access
- Lustre – HSM and SLURM
  - Data storage and processing
- ArmadilloDB
  - Data index and access

Platform overview



Compute nodes

+

Tier0

memory space

Service nodes

- Lustre
- Mgmt. node with job scheduler

Cloud access

Login node

Cloud FE (ProActive)

NoSQL (ArmaDB)

Tier1

online storage

Tier2

nearline storage

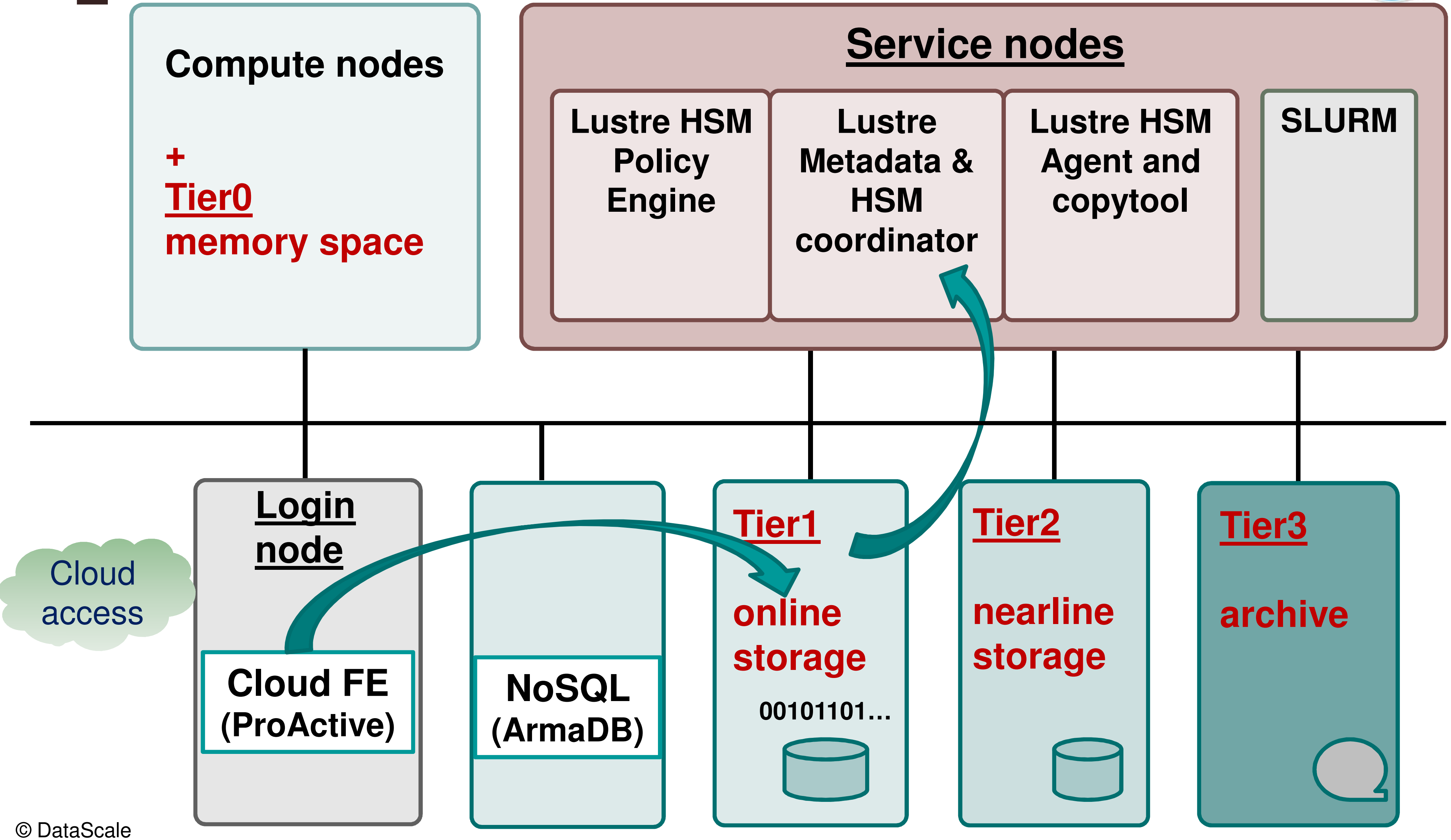
Tier3

archive

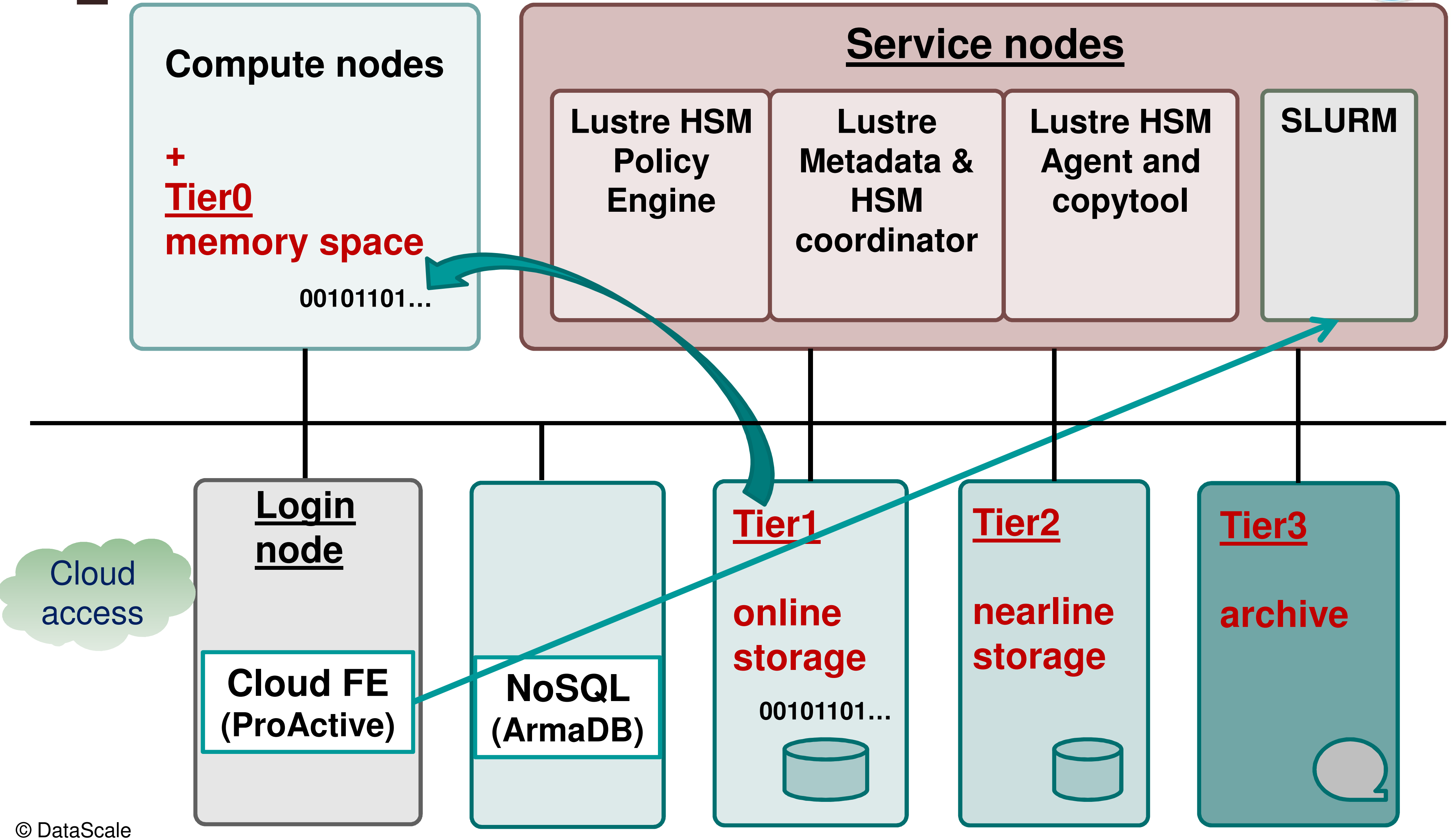
Three key mechanisms

1. Cloud Interconnect
   - Input data from the cloud, output result to the cloud
2. Optimized Data Management
   - Hierarchical Storage Management
   - Automatic data movements
3. NoSQL database and data mining
   - Distributed database

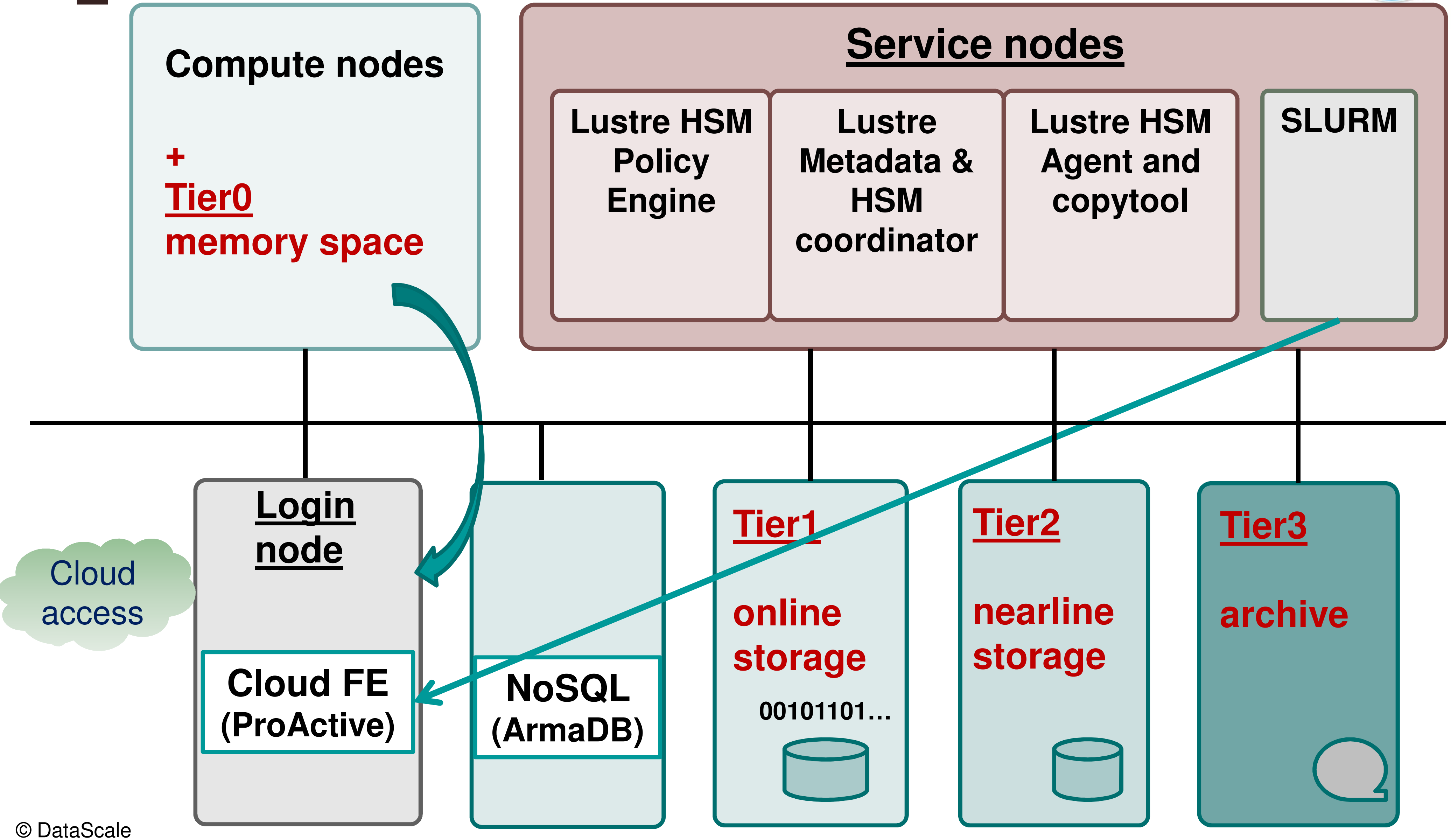
Cloud Interco
Data input



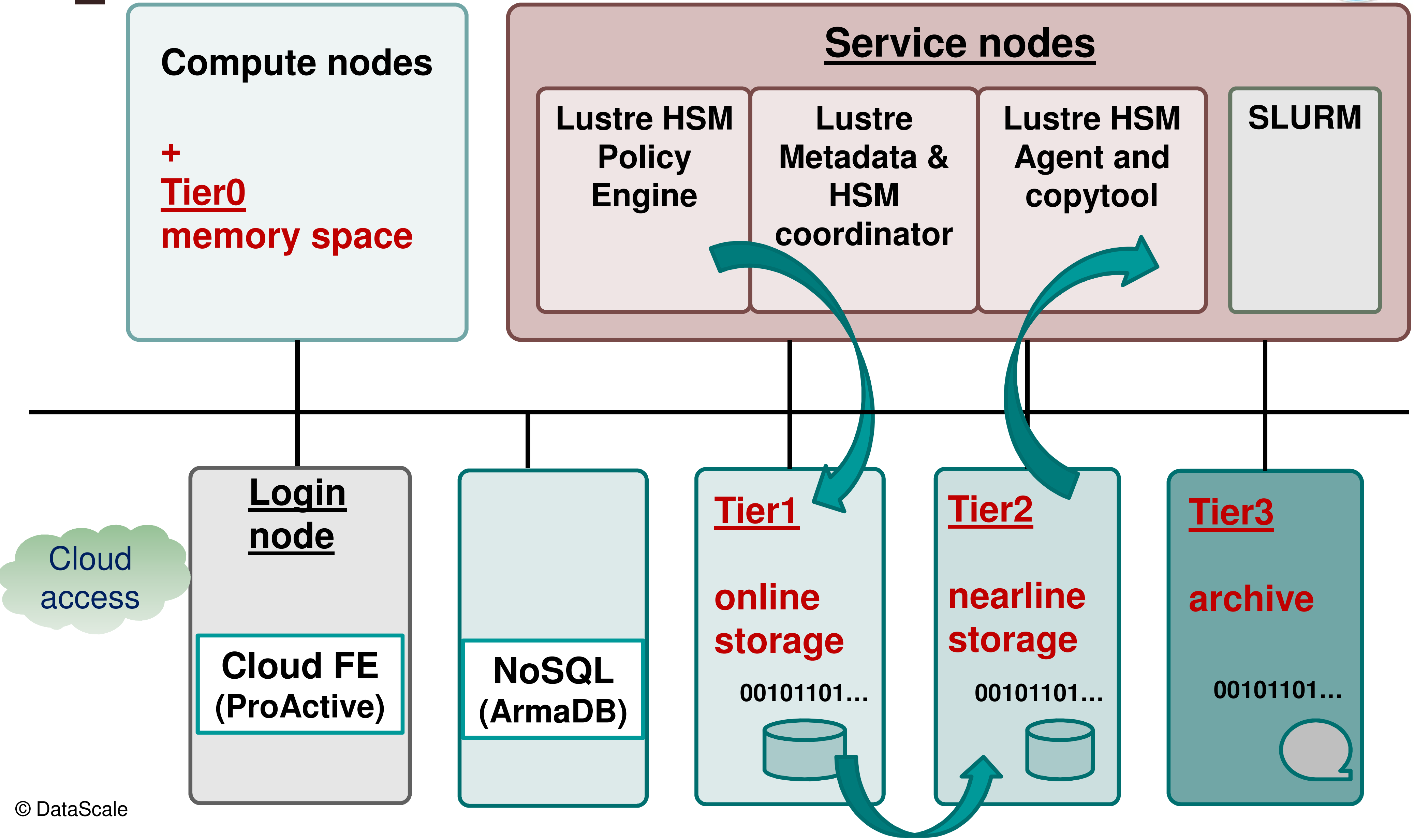
Cloud Interco
SLURM activation



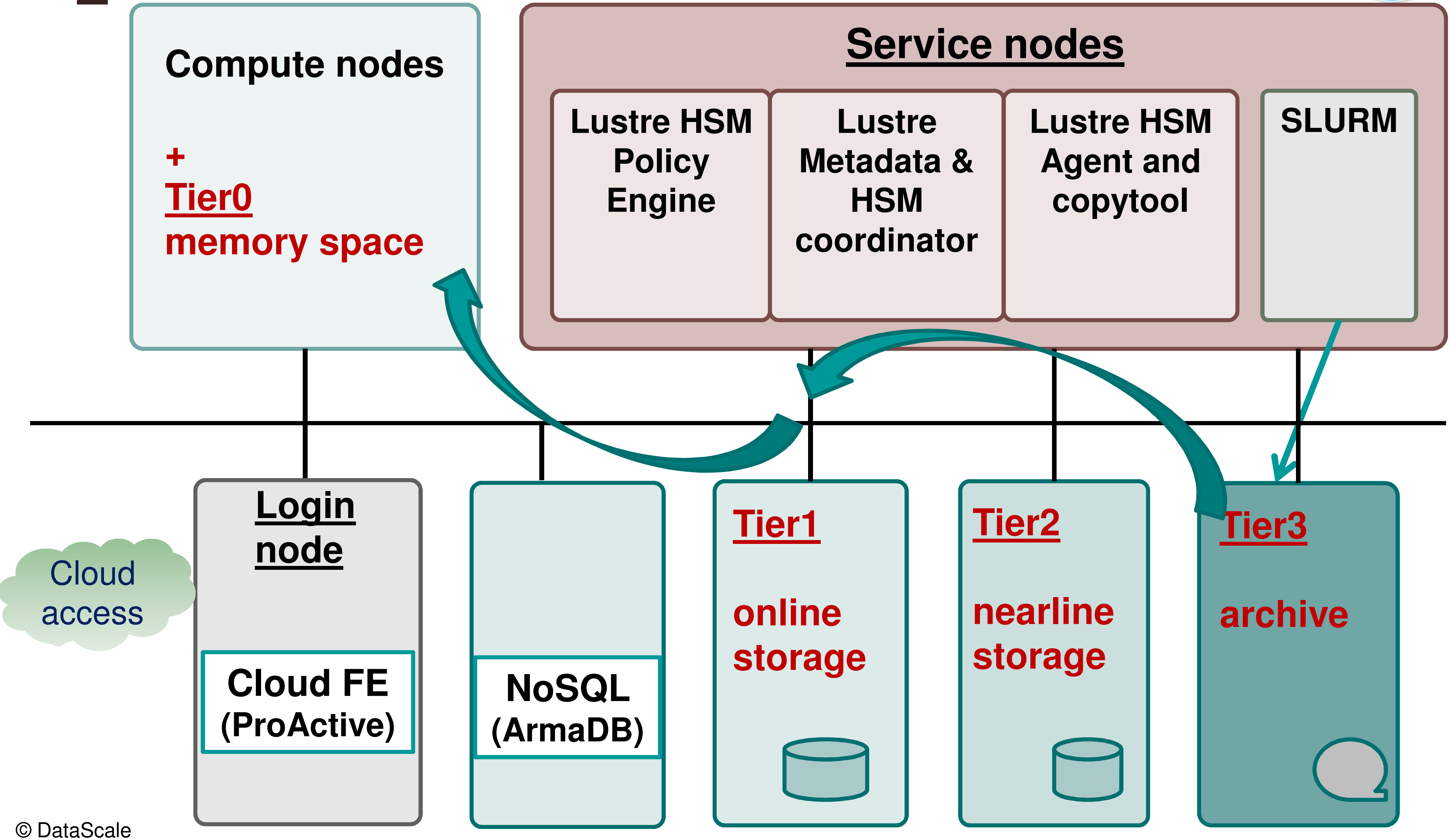
Cloud Interco
Results export



Data management/HSM
Lower storage cost



Data management/HSM
Data preload



Compute nodes

+ Tier0 memory space

Service nodes

Lustre HSM Policy Engine

Lustre Metadata & HSM coordinator

Lustre HSM Agent and copytool

SLURM

Cloud access

Login node

Cloud FE (ProActive)

NoSQL (ArmaDB)

Tier1 online storage

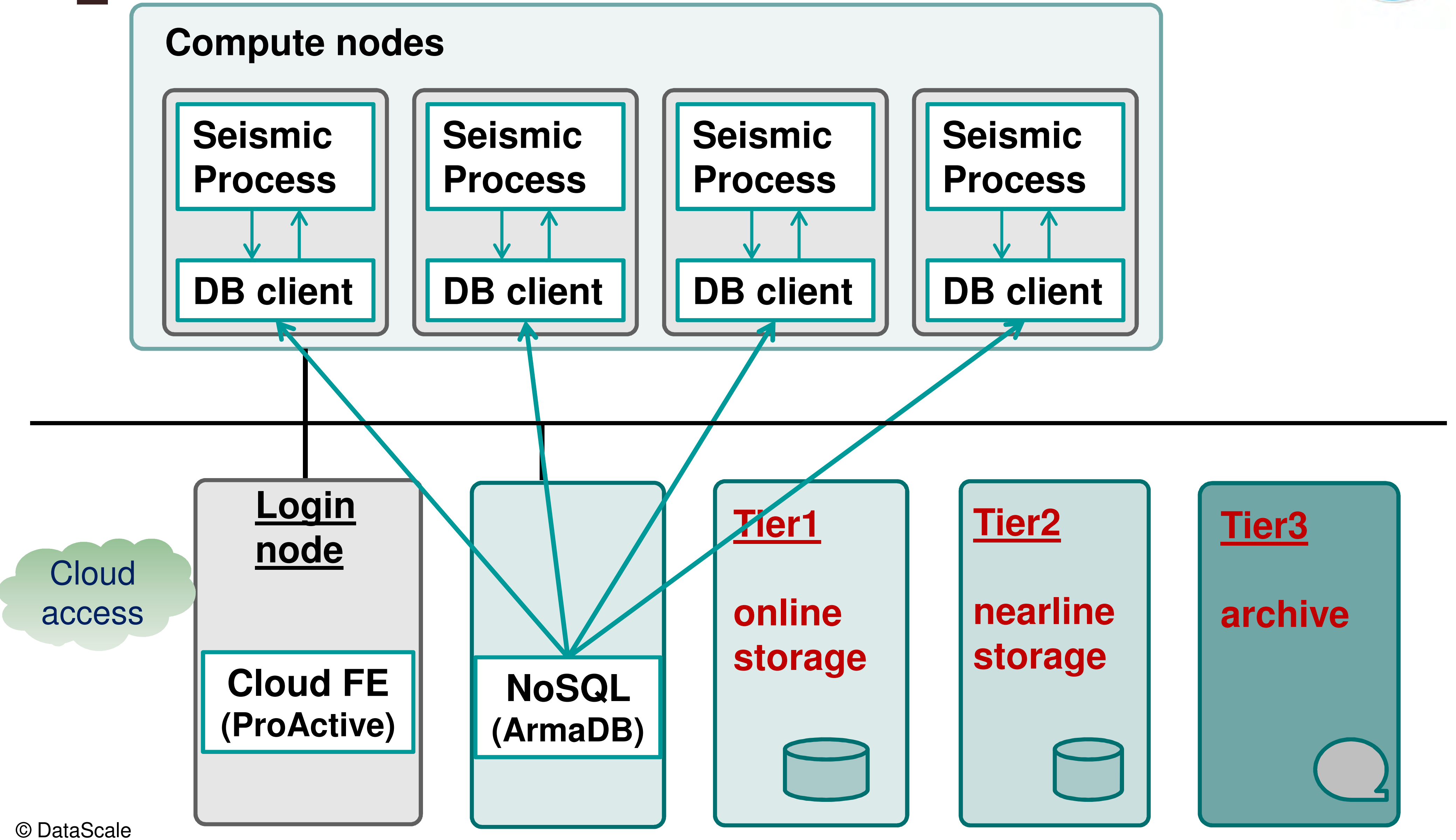
Tier2 nearline storage

Tier3 archive

NoSQL Database

- Filter data prior to mining
- SQL/NoSQL Hybrid
  - list, map, etc.
- Automatic Scaling
  - Master on Login Node
  - Replicated on compute nodes
  - Writes buffered on nodes, batch writes on master

NoSQL Database



Conclusion

- Experimentation phase
- Expected results
  - Seismic event detection
    - Real-time monitoring
    - 10-years analysis
  - Cluster Management
    - In-situ extraction algorithm, implemented on the cluster's nodes, for failure detection
  - Multimedia product analysis
    - Real-time image correlation

Thank you