Plateformes big data
Quels besoins pour quelles utilisations ?

Participants

- **OSIRIM**, Mohand Boughanem, IRIT, Toulouse
- **GALACTICA,** Farouk Toumani, LIMOS, Clermont-Ferrand
- **PAGoDA,** Emmanuel Coquery, LIRIS, Lyon
- **PerSCiDO,** Marie-Christine ROUSSET, LIG, Grenoble

Plateformes big data
Quels besoins pour quelles utilisations ?

- Appels à projets soutien aux Plateformes de l'INS2I
- Journées Plateformes, 6-7 octobre 2016 à Clermont-Ferrand (https://indico.in2p3.fr/event/13365/)
  - Organisation : PlaScido, ANR MDK et GDR MaDICS,
  - 4 sessions : PlaScido, Plateforme de gestion de données scientifiques, Datacenters, Big Data & HPC
- **Organisation d'un atelier plateformes en 2018** (en liaison avec le GDR MaDICS)

Plateformes big data
Quels besoins pour quelles utilisations ?

- Qu'est-ce qu'une plateforme ?
- Quels besoins ? Quels objectifs ?
- Quels types d'utilisateurs ? Quels services ?
- Quelles modalités d'usage ? Quel degré d'ouverture ?
- Premiers retours d'expériences
- Quel positionnement dans le paysage national/offres privées ?
- Quelles sont les évolutions envisagées ?
    - Réseaux de plateformes ?

Plateformes big data
Quelques problématiques générales

- Reproductibilité des expérimentations
- Données sensibles/propriétés des données/vie privée
- Partage de gros volumes de données a des fins d'expérimentation
- . . .

OSIRIM

Observatoire des Systèmes d'Indexation et de Recherche d'Information Multimédia

Plateforme pour l'exploitation de grands volumes de données

Mohand Boughanem, Philippe Joly, Guillaume Dubreule, Jacques Thomazeau

IRIT — CNRS — Université Toulouse III Paul Sabatier — INP Toulouse — Université Toulouse Jean Jaurès

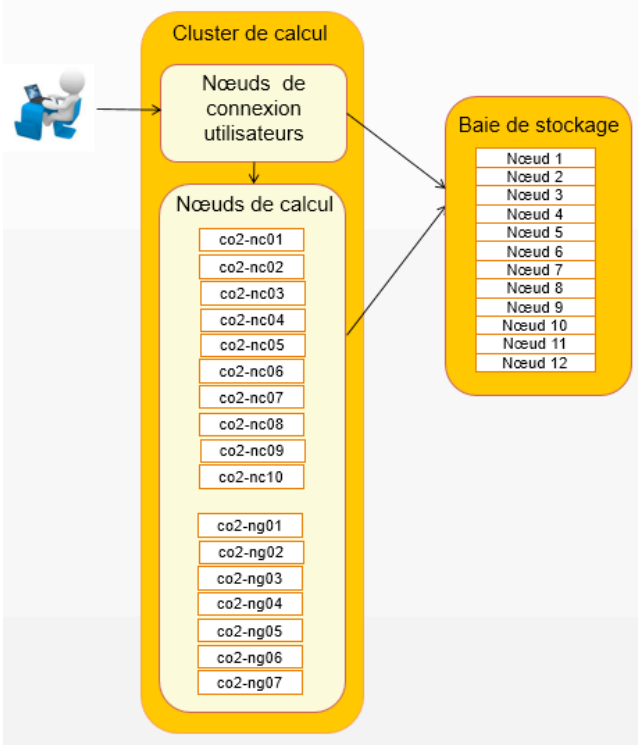
Observatoire des Systèmes d'Indexation et de Recherche d'Information Multimédia

- Plateforme matérielle localisée à et administrée par l'IRIT, réalisée dans le cadre du Contrat de Plan Etat Région (CPER) 2007-2013,

- Ouverte aux chercheurs et étudiants travaillant sur des sujets liés au traitement de grands volumes de données, ainsi qu'à la communauté informatique et autres domaines scientifiques souhaitant utiliser ses moyens matériels ou logiciels

Offre de services

- Un cluster de calcul ….
  - 10 serveurs IBM équipés chacun de 512 Go RAM et 64 cœurs
  - + 7 serveurs Dell équipés chacun de 4 cartes GPU Nvidia
  - Une solution de stockage Isilon 1 PO (12 nœuds de 36 disques de 3 To chacun)
- … servant de support à une offre de services logiciels mutualisés :
  - Un gestionnaire de jobs et de ressources SLURM pour la distribution de traitements réalisés avec des langages / logiciels mutualisés : C++, PYTHON, JAVA, R, …
  - Un cluster HADOOP (Hortonworks) avec son écosystème applicatif : SPARK, HIVE, HBASE, …
  - Un cluster de calcul GPU support à l'utilisation de frameworks de Deep Learning (TensorFlow, Pytorch, Theano, Keras)



Objectifs

- Héberger des projets scientifiques nécessitant :
  - le stockage et
  - le partage de plusieurs téraoctets de données

→ Réaliser des expérimentations sur de grands volumes.

- Partager/publier des données d'expérimentation:
  - Exemple : 1% des tweets mondiaux (streaming) depuis septembre 2015, Corpus TREC, ...

- Partager des outils, vos outils, logiciels utilisés dans vos expérimentations)
  - Hadoop, Terrier, Deep Learning, …

→ Permettre la reproductibilité des résultats.

Modalités d'usage d'Osirim

- OSIRIM est ouverte
  - Aux chercheurs et étudiants travaillant sur des sujets liés au traitement de grands volumes de données.
  - À la communauté informatique et autres domaines scientifiques souhaitant utiliser ses moyens matériels ou logiciels sous certaines conditions.
- … pour l'hébergement de projets
  - Un projet est un espace d'hébergement de données et de logiciels partagés par plusieurs utilisateurs. Il est placé sous la responsabilité d'une personne.
- Comment faire héberger un projet sur OSIRIM
  - Soumettre une demande d'hébergement via le site web «http://osirim.irit.fr», examinée par un comité de pilotage mensuel.
  - Accepter la charte d'utilisation de la plateforme.

Utilisateurs

- Travaux de recherche des équipes de l'IRIT
  - 100 utilisateurs

- Projets (multipartenaires)
  - QUAERO (terminé) : innovation sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues (IRIT, IRISA, Exalead (Dassault)).
  - SemDis : création de bases distributionnelles de référence pour le français.
  - CAIR : recherche agrégative de données (IRIT, LIRIS).
  - POLEMIC: analyse du comportement des utilisateurs dans les réseaux sociaux (IRIT, UAM Mexico).
  - COMPUBIOMED : Meta mining pour la recommandation en biosanté (IRIT, INSERM).
  - LISTIC : Réseaux sociaux numériques aux réseaux sociaux reels (IRIT, LISTIC) .
  - …
- Ouverte à tous les chercheurs (modalités d'usage)
  - UMR LISIS (Univ. Marne la vallée)
  - CLEE (Univ. Jean Jaurès)
  - Petasky (terminé) : techniques de partitionnement de données (LIRIS).
  - …

Mais aussi …

- Participations à des campagnes d'évaluation
  - TREC (Text Retrieval Conference), INEX (XML Retrieval), CLEF (Cross Language Evaluation Forum), TrecVid (TREC Video Retrieval Evaluation), mais aussi OAEI (Ontology Alignment Evaluation Initiative).

- Soutien pour l'initiation à la recherche dans des formations de master
  - Master SID Université Toulouse 3 : apprentissage de technologies Hadoop (Hive).
  - Master M2 IT/ Enseeiht : Fouille de tweets.

- Accompagnement d'évènements spécifiques
  - Hackday CORIA/CIFED 2016.

Perspectives d'évolution

- Compléter Osirim avec une approche de type « cloud » proposant des services sur étagères (bases de données NoSql, …) ou la mise à disposition de machines virtuelles dédiées

- Renouvellement de l'infrastructure matérielle

- Mise en réseau avec les autres plateformes

Fin

Pour tout contact et demande d’hébergement :

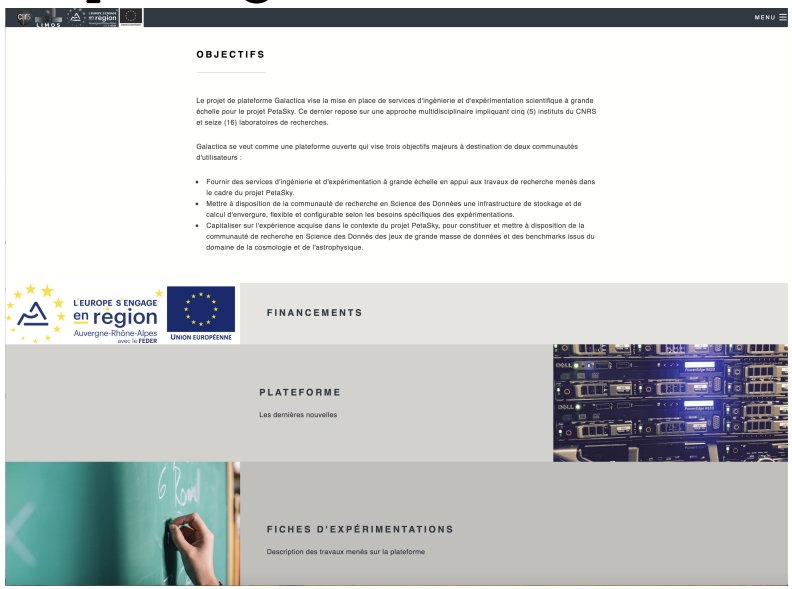
http://osirim.irit.fr

osirim@irit.fr

La plateforme Galactica

F. Toumani, LIMOS, CNRS, UCA

https://galactica.isima.fr



- A l'origine ..
  - Appel à projets *"Soutien Plateformes Science des Données"* (PlaSciDo), 2015 de l'INS2I
  - Emanation des travaux de recherche du projet PetaSky (2012-2016), programme Mastodons
  - Partenaires : LIMOS, LIRIS, LAL, LIF, LABRI, PRISM, OBSPM, LAM, APC, CC IN2P3
  - Financement : l'INS2I et Contrat Plan Etat Région (CPER) de la région Auvergne
  - Administrateur : Frédéric Gaudet (gaudet@isima.fr)

Objectif : *Une plateforme ouverte qui vise le développement et la mise à disposition de services d'ingénierie et d'expérimentation à grande échelle pour les chercheurs dans le domaine des grandes masses de données*

- Un cluster composé de noeuds de calcul et de noeuds de stockage
- Une puissance de calcul totale de 128 coeurs (256 vCPU) à 2,40 GHz, 3,8 To de RAM et une capacité de stockage de 143 To
- Un réseau de 10 et 40 Go/s

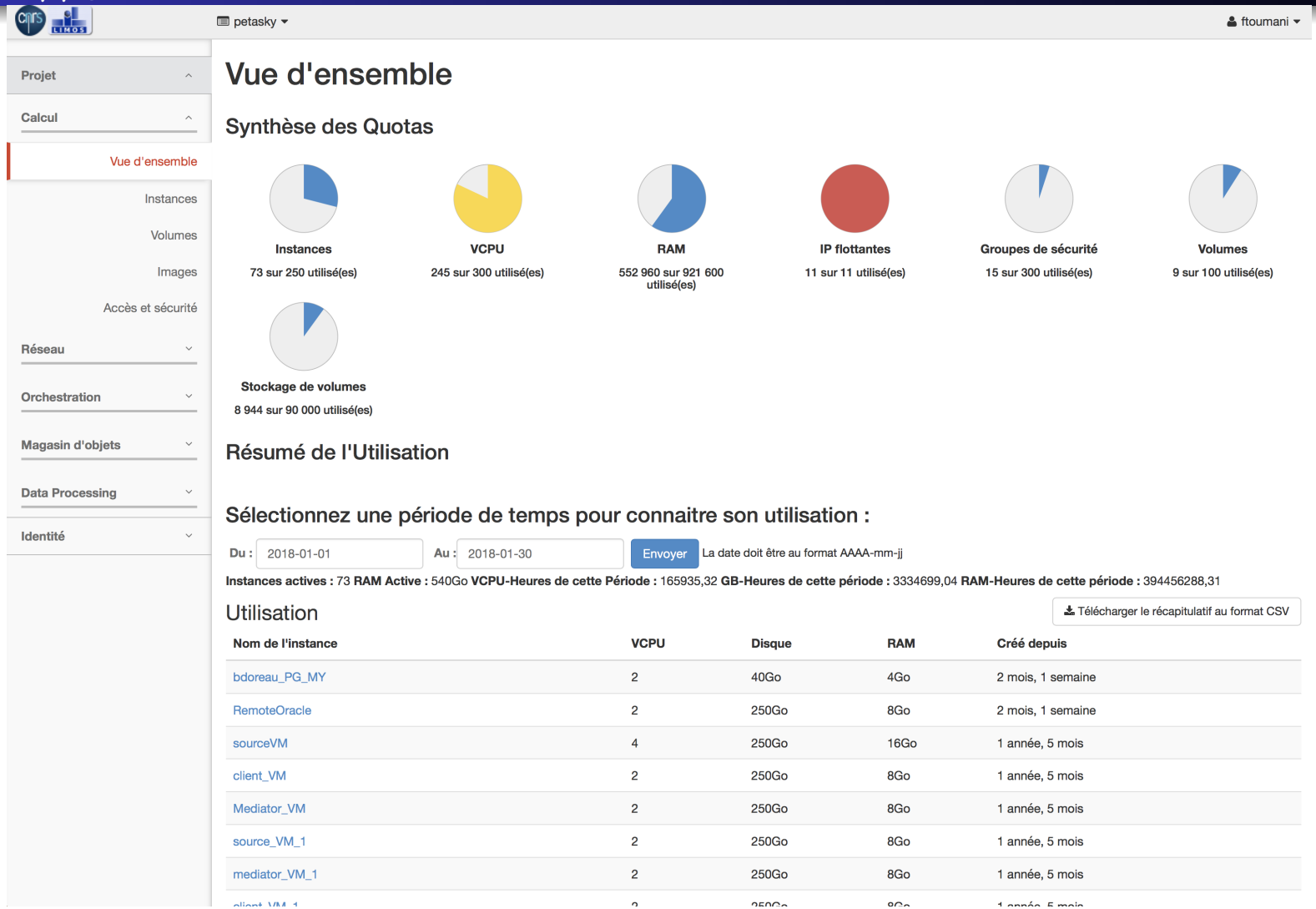
Une infrastructure de calcul et de stockage flexible et facile à configurer pour s'adapter aux besoins spécifiques des expérimentations dans le domaine des grandes masses de données

Services GALACTICA

1. Infrastructure en tant que service
   - Création par un chercheur (de manière aisée), au sein d'un environnement virtualisé, des ressources informatiques (calcul, stockage, réseau, etc.) adaptées à ses besoins
2. Création par un chercheur d'un cluster de traitement de données à partir de modèles prédéfinis (templates)
   - Exemples de modèles : Apache Hive, Apache Hadoop, Hortonworks Data Platform, Cloudera et Apache Spark
   - Création et exécution de jobs sur des clusters déployés au sein de la plateforme (Elastic Data Processing)
3. Mise à disposition de jeux de données
   - Sous forme brute : LSST (2 To et 35 To), GAIA DR1 (730 Go compressés) et SDSS DR9 (8 To, http://www.sdss3.org/dr9)
   - Sous forme de source de données : une collection de données + une infrastructure logicielle permettant d'exploiter ces données
     Exemple : VM + SGBD MySQL + BD SDSS DR9

La plateforme GALACTICA

https://galactica.isima.fr



- Formation assurée par Frédéric Gaudet
- Programme de formation
  - Présentation de l'architecture Galactica (computes, stockage Ceph..)
  - Gestion des instances
  - Gestion des volumes
  - Utilisation du stockage objet
  - Utilisation de l'orchestration
  - Utilisation de l'Elastic Data Processing (Hadoop & Spark)
  - Règles de sécurité
- Sessions de formation régulières et à la demande

⇒ GALACTICA Aujourd'hui

- 19 projets, une soixantaine d'utilisateurs, 12 laboratoires de recherche
- Comment utiliser Galactica ?
  Transmettre une fiche d'expérimentation à Frédéric Gaudet

⇒ Evolution

- Développement de nouveaux services : *métrologie* pour les expérimentations
- Augmentation de la capacité de calcul et de stockage
- Mise en réseau avec d'autres plateformes
  - Pagoda
    - Réflexion commune sur le choix d'infrastructure
    - Développement concerté et complémentaire des expertises
    - Intégration ou fédération ?
  - PerSCido/OSIRIM

LIRIS

UMR 5205 CNRS

PAGoDA

Plateforme à base de plugins pour les Application biG Data Analytics

Laboratoire d'InfoRmatique en Image et Systèmes d'information

cnrs INSA Lyon 1 UNIVERSITÉ LUMIÈRE LYON 2 ÉCOLE CENTRALE LYON

Exemples

PetaSky

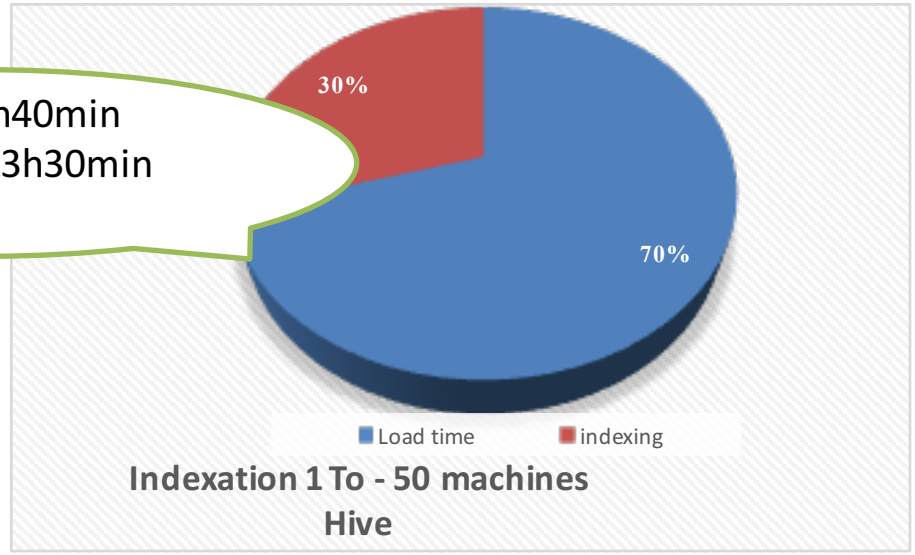
- Grosses volumétries
- Temps de traitements élevés

CAIR

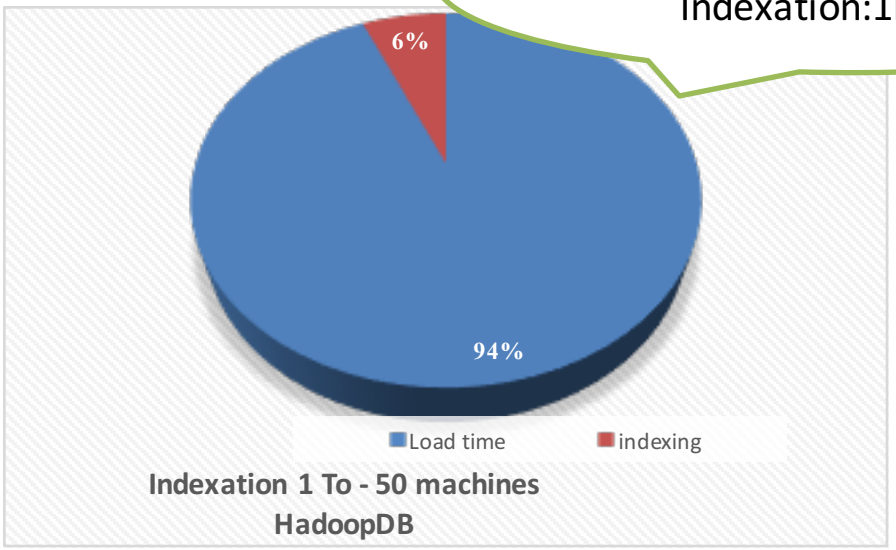
- Reproduction d'expérimentations
- Réutilisation de dispositifs (logiciels) expérimentaux

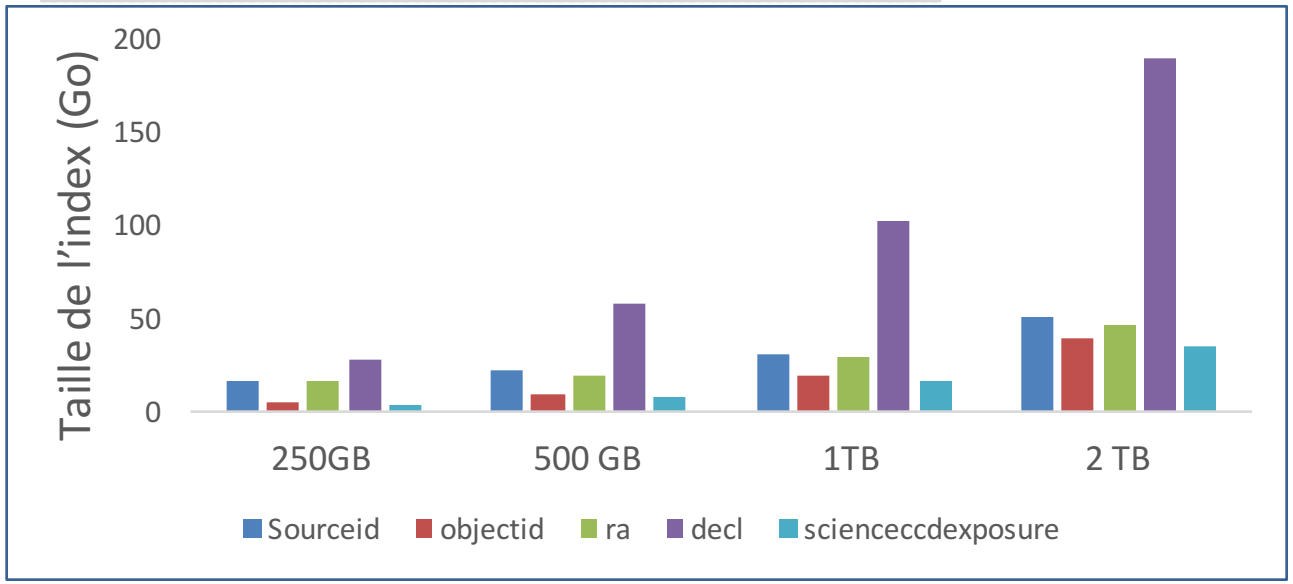
Chargement de données

Total: 11h40min
Indexation:3h30min



Total:23h20min
Indexation:1h30min

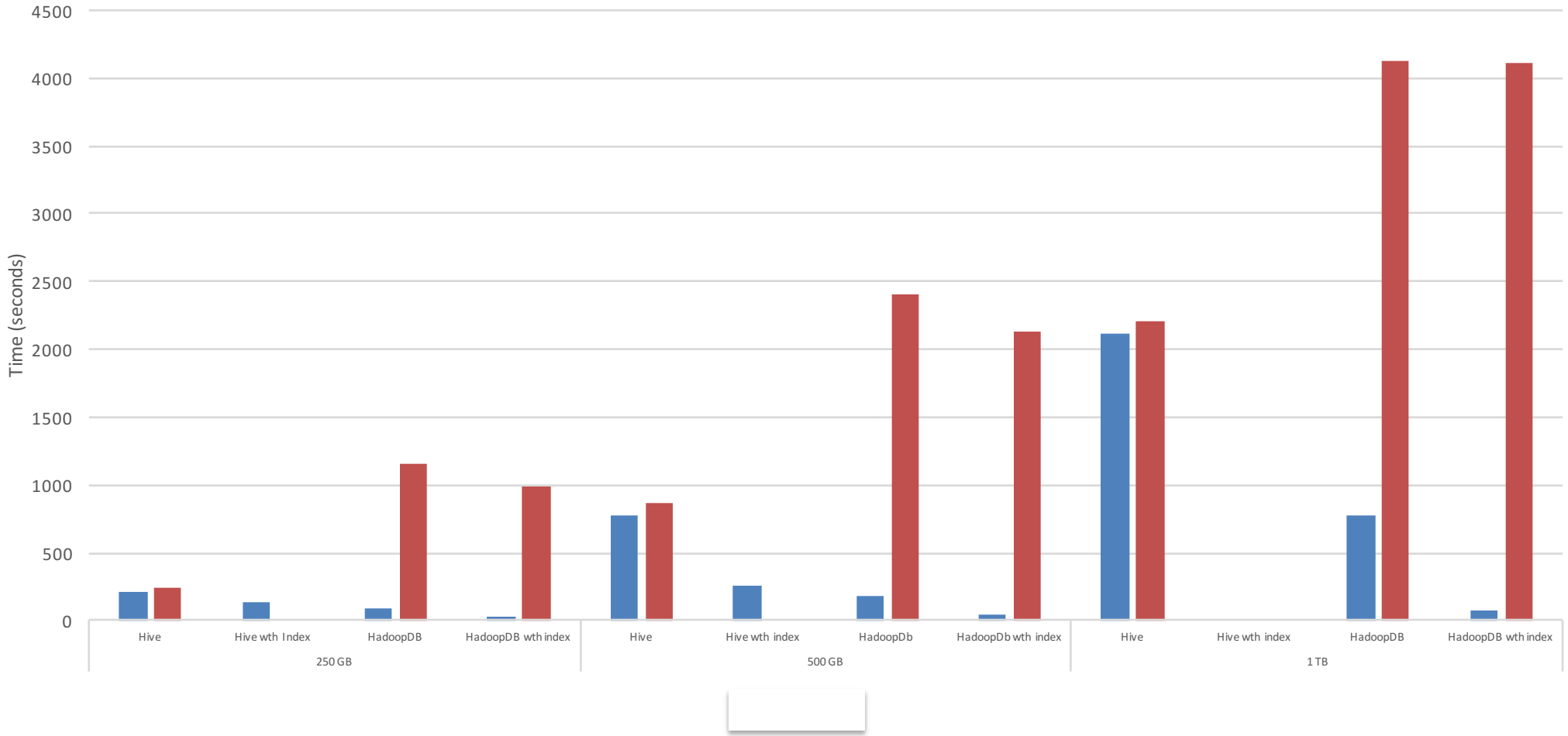




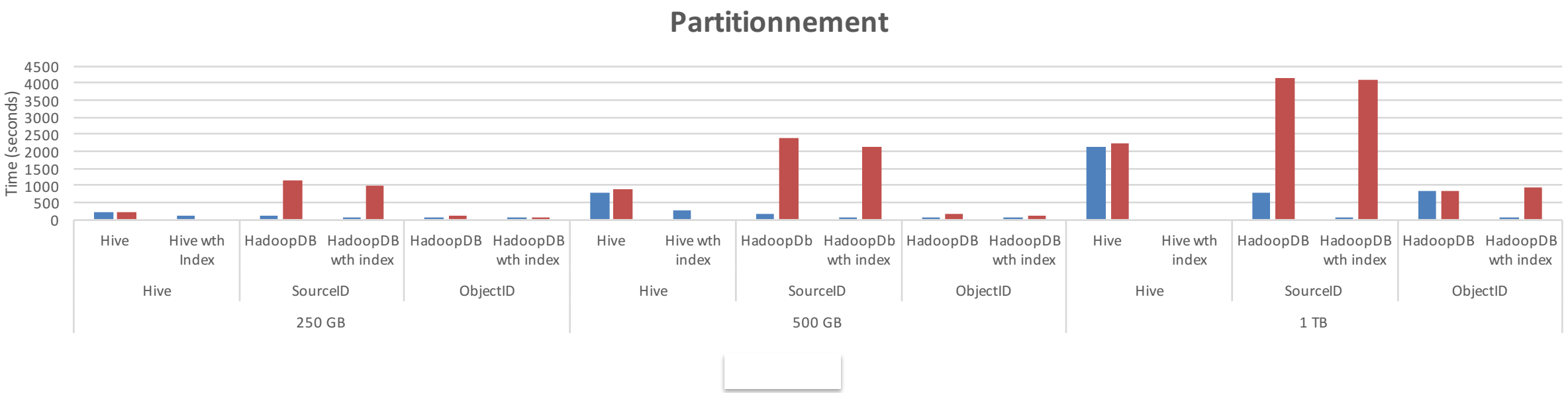
- Hive:

  25 >> 50 machines: gain de 15%

Traitement de requêtes



Traitement de requêtes



CAIR

Contradictions dans l'état de l'art des performances des algorithmes les plus utilisés (Gaston, FFSM, gSpan)

Gaston est l'algorithme le plus rapide parmi FFSM, gSpan, MoFa (Wörlein et al., 2005)

Gaston est le dernier en terme de temps d'exécution comparé à FFSM, FSP, SPIN et CloseGraph (Rehman et al., 2014)

FFSM est plus performant que gSpan (Huan et al., 2003)
FFSM a des performances considérables par rapport à gSpan (Patel et al., 2013)

gSpan est un peu plus rapide que FFSM (Wörlein et al., 2005) (Meinl et al., 2006)
gSpan est le meilleur algorithme en terme de consommation de mémoire comparé à FFSM, MoFa, Gaston (Wörlein et al., 2005) (Meinl et al., 2006)
gSpan est aussi compétitif que Gaston et FFSM avec des fragments pas très grands (Douar et al., 2014)

gSpan est beaucoup plus performant que FSG (Yan et al., 2002)
AcGM est plus rapide que FSG (Inokuchi et al., 2002)

gSpan et FSG sont les algorithmes de fouille de graphes les plus performants dans leurs catégories respectives (Douar et al., 2014)

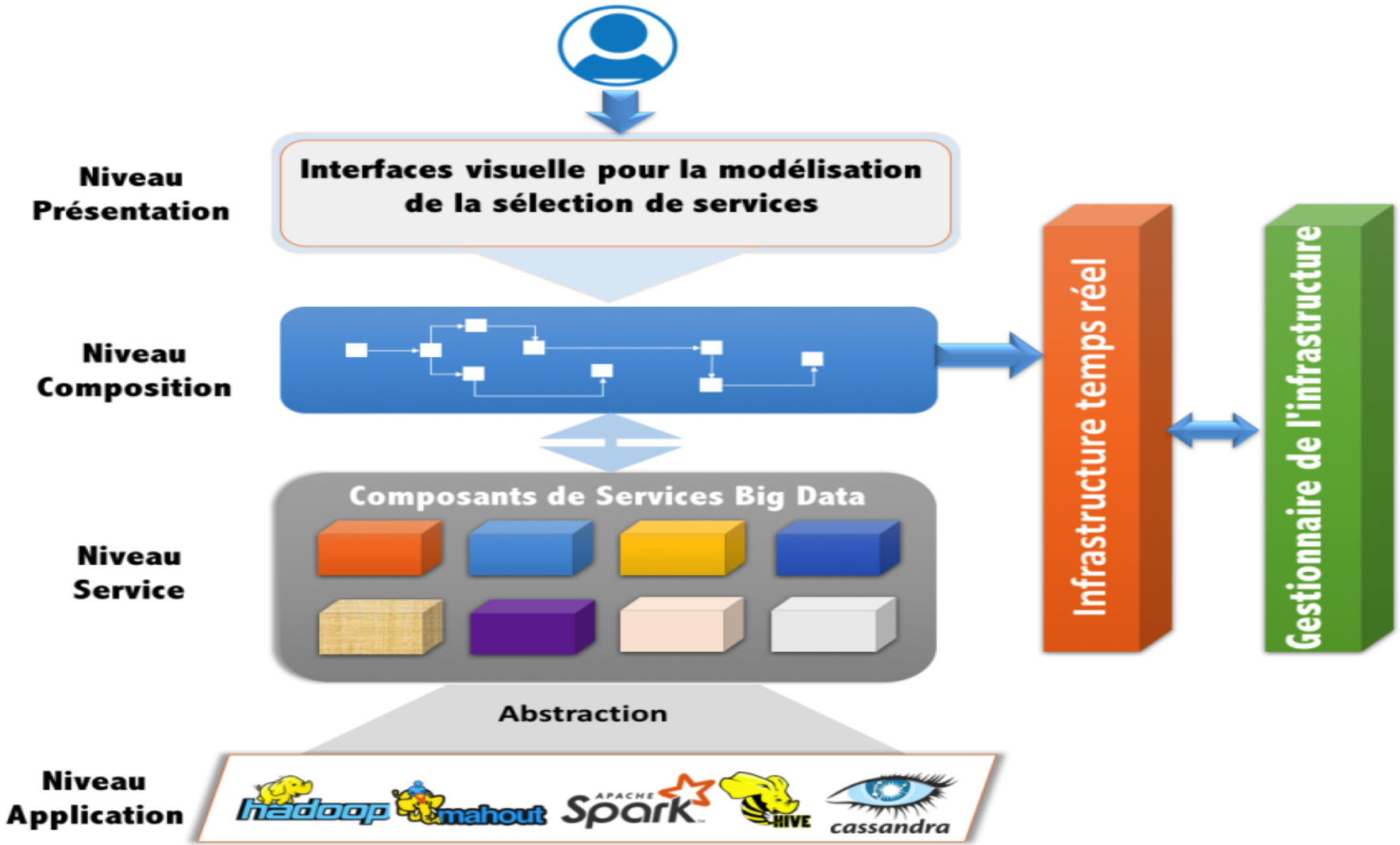
Constat

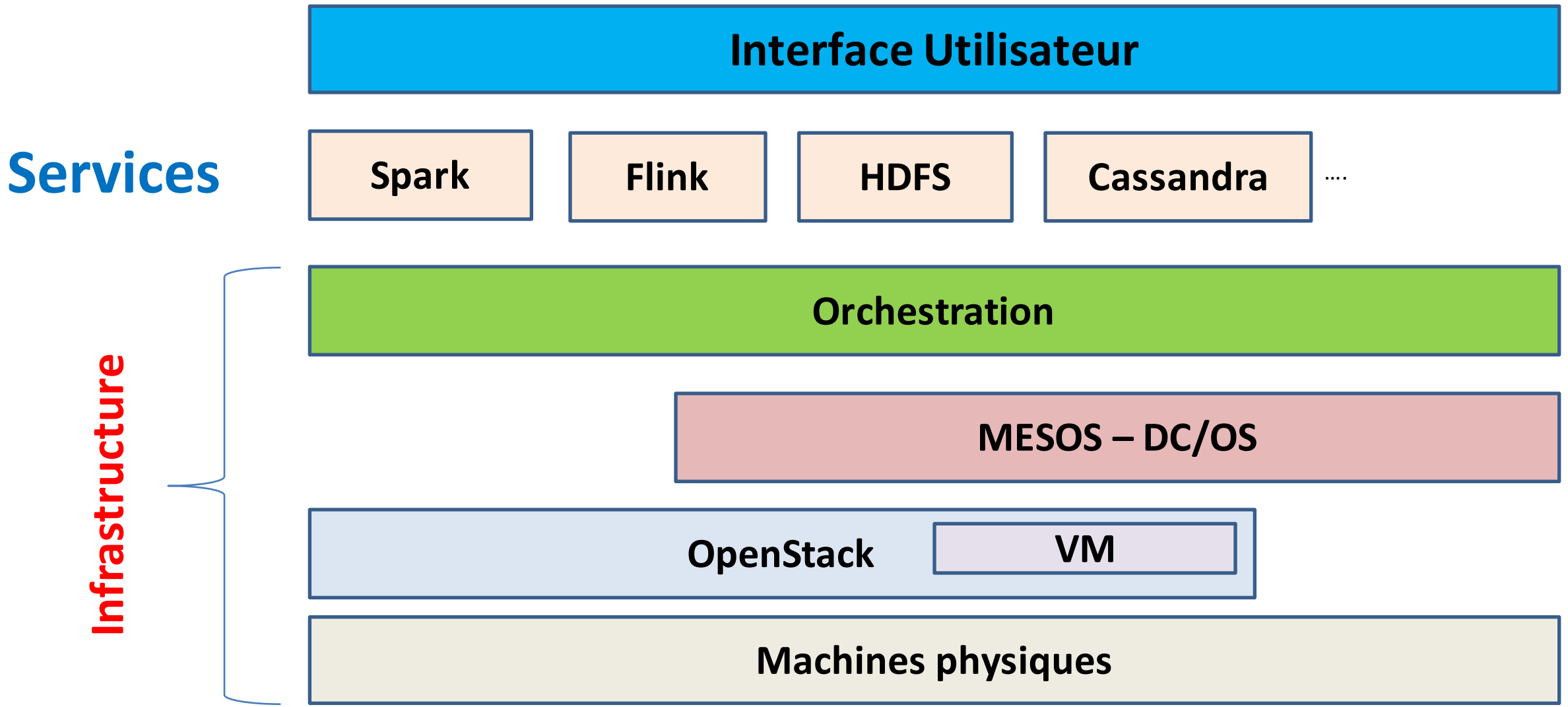
- Configuration et déploiement d'applications :
  - Spécification d'une chaîne de traitements complexes
  - Le déploiement

Processus complexe (investissement technique)

- Identification des composants et des sources
- Composition des composants
- Dépendances entre composants

- Déploiement d'infrastructures virtuelles (clusters de machines virtuelles ou des containeurs (*ex.*, *Docker*)) ➔ **conserver et reproduire aisément des contextes expérimentaux**
- Déploiement de clusters de machines préconfigurées (par exemple pour les flux de données)
- Assembler aisément plusieurs briques de calcul (*ex., Openshift*, *Knime*)
- Collecte et analyse de logs et autres mesures expérimentales (*ex., Elastic Stack*) couplé à un ou plusieurs systèmes d'analyse de données


Niveau Présentation
Interfaces visuelle pour la modélisation de la sélection de services
Niveau Composition
Niveau Service
Composants de Services Big Data
Abstraction
Niveau Application
hadoop
mahout
APACHE Spark
HIVE
cassandra
Infrastructure temps réel
Gestionnaire de l'infrastructure


Interface Utilisateur
Services
Spark
Flink
HDFS
Cassandra
....
Orchestration
Infrastructure
MESOS – DC/OS
OpenStack
VM
Machines physiques

PerSCiDO_Grenoble_Alpes :
Principes et fonctionnalités d'une plateforme ouverte et interopérable de partage de jeux de données

https://persyval-platform.univ-grenoble-alpes.fr/

Responsable: Marie-Christine ROUSSET

Lucie Albaret[(5)], Brigitte Bidegaray[(1,2,4)], Pierre Hébert[(2)], Fabrice Jouanot[(3)], Alireza Moussaei[(1)]

Université Grenoble Alpes et CNRS

(1) CNRS, (2) Labex PERSYVAL-lab, (3) LIG, (4) LJK,

(5) Service inter-établissement de la Documentation Université Grenoble Alpes - Grenoble INP

Principes

- Découpler le stockage des jeux de données de leur description par des **méta-données** riches, flexibles et évolutives
  - interrogation avancée de ces méta-données
  - référencement de jeux de données pouvant être stockés sur différents serveurs ou data centers
- Faciliter l'interopérabilité avec d'autres plateformes
  - e.g., DataCite, Osirim, Galactica
- Suivre les standards
  - du Linked Open Data en termes de modèle de données (RDF) et aussi de vocabulaires spécialisés de métadonnées comme Dublin Core, Friend Of a Friend, Creative Commons, etc
  - émergents de consortiums internationaux comme DataCite, FaBio, Radar, etc … qui visent la définition de standards pour citer et décrire des données de recherche.
- Anticiper les usages et inciter aux bonnes pratiques

Inciter aux bonnes pratiques

- Pousser les chercheurs à référencer leurs jeux de données par des **identifiants externes persistants** (HAL, DOI, etc …)
  - Un DOI (Digital Object Identifier), prôné par DataCite

  => Convention signée avec l'INIST (représentant français de DataCite) pour que PERSYVAL-lab puisse délivrer des DOIs
- Pousser les chercheurs à anticiper **la citation** souhaitée pour leur jeu de données

  => champ pré-rempli en cas d'un DOI existant
- Pousser les chercheurs à **préciser le droit d'usage** de leurs jeux de données **par une licence Creative Commons**

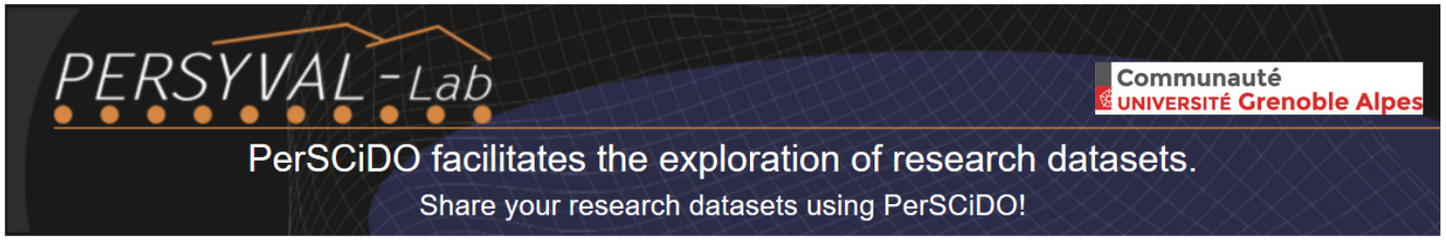
  ⇒ Menu déroulant avec les différentes licences fournies sous la forme d'un vocabulaire contrôlé
- **Eviter tant que possible la saisie de chaines de caractères « libres »** pour remplir les valeurs de champs à renseigner

  => Menus déroulants avec des valeurs prédéfinies (des constantes dans la BD)

Plateforme ouverte mais sous contrôle

- Interface accessible à tous:
  - pas de restriction pour naviguer et rechercher des jeux de données par leurs descriptions
    - login obligatoire au clic de demande de téléchargement d ’un jeu de données
    - + si jeu de données déclaré en accès restreint, envoi automatique d’un email de demande d’autorisation au chercheur qui a déposé ce jeu de données
  - login obligatoire pour soumettre un nouveau jeu de données
    - le chercheur décrit lui-même les méta-données de son jeu de données via l’interface
      - à la fin, Il a le choix de demander le dépôt sur PerSciDo ou de fournir une URL où trouver son jeu de données
    - validation (en ligne) de la mise en ligne via PerSciDo par un comité éditorial

PerSCiDO



Numbers

Datasets: 25
Downloaded: 188

Explore PerSCiDO research data collections and related publications

Search terms... search

Submit a new dataset

Recent datasets

Recently Published | By Scientific Field | By Data Type

2017 Dec 18 Restricted Survey data

Professional SNA

Depositor: Aria Teimourzadeh

This dataset contains the experience of 184 social network users for job seeking and hiring purposes in France.

2017 Dec 15 Open Experimental data

F-TRACT, ATLAS Decembre 2017

Depositor: Olivier David

Connectivity probability with associated p-values as well as features describing fibers biophysical properties, estimated from CCEP data recorded in 213 patients, in the MarsAtlas, Brodmann, AAL and MaxProbMap parcellation schemes. The CCEP features are: peak and onset latency (LatStart), amplitude, integral, duration and the velocity estimated from the onset latency and the fibers distance between the parcels. Features maps : Images representing the connectivity probablility and response features for all the regions in the MarsAtlas parcellation.

2017 Dec 15 Open Speech data

Translation Augmented LibriSpeech Corpus

Depositor: Laurent Besacier

Large scale (>200h) and publicly available read audio book corpus. This corpus is an augmentation of LibriSpeech ASR Corpus (1000h) and contains English utterances (from audiobooks) automatically aligned with French text. Our dataset offers ~236h of speech aligned to translated text. Speech recordings and source texts are originally from Gutenberg Project, which is a digital library of public domain books read by volunteers. Our augmentation of LibriSpeech is straightforward: we automatically aligned e-books in a foreign language (French) with English utterances of LibriSpeech. We gathered open domain e-books in French and extracted individual chapters available in LibriSpeech Corpus. Furthermore, we aligned chapters in French with English

News

Start of the collaboration with the GALACTICA platform
10/11/2017

New version of Perscido Platform
08/11/2017

Start of the collaboration with the OSIRIM platform
08/11/2017

Presentation of PerSCiDO at data4ist day (Paris)
03/06/2016

Requesting data isn't harassment, and refusing to share data isn't science

Une page d'une des étapes de soumission

PerSCiDO

New dataset About Support Admin Si

1. General Information | 2. Content Description | 3. Datatype Content | 4. Data Access

Please describe your dataset in as much detail as possible. A detailed description will make it easier for others to find your dataset in PerSCiDO.

Fields marked with an asterisk (*) are required. For more information on expected content for a field, mouse over the ? icon.

Please select the data type of your dataset

Trace data | Execution Trace

If your dataset has been processed for an automatic task,

please select the corresponding task(s) below

- Anomaly detection
- Classification
- Clustering
- Dimension Reduction
- Grammatical inference
- Pattern extraction
- Prediction
- Preference learning
- Regression Analysis
- Rule extraction
- Visualisation

Next

PowerdBy ... - Term of service - Contact us

Evolution

- Mise en réseau avec les plateformes partenaires
  - Consolider et étendre la collaboration avec les plateformes Osirim, Galactica et Pagoda

  => Aider les chercheurs à trouver la plateforme qui répond le mieux à leurs besoins

- Etendre PerSciDo à la description d'algorithmes, de services et de chaines de traitement de données,
  - Nouvelles méta-données
  - Liage entre plateformes, jeux de données et traitements

  ⇒ Offrir un support pour guider la rédaction et le suivi de DMPs

⇒ **créer un réseau de plateformes pour une science ouverte et des traitements de données reproductibles**