À la poursuite des Big Data

« Big Data 2.0 »

Mercredi 27 mars 2013

Coordination scientifique :
– ***Jean-Michel Batto*** *(INRA)*
– ***Pierre Léonard*** *(INRA)*

Amphithéâtre Becquerel, École Polytechnique, Palaiseau

http://www.association-aristote.fr
info@association-aristote.fr

Edition du 3 prairial an CCXXI (*vulg.* 22 mai 2013)

ARISTOTE Association Loi de 1901. Siège social : CEA-DSI CEN Saclay
Bât. 474 91191 Gif-sur-Yvette Cedex.
Secrétariat : Aristote, École Polytechnique, 91128 Palaiseau Cedex.
Tél. : +33(0)1 69 33 99 66 Fax : +33(0)1 69 33 99 67
Courriel : Marie.Tetard@polytechnique.edu
Site internet http://www.association-aristote.fr

Table des matières

Chapitre 1

Programme de la journée

1.1 Introduction

Presque deux ans après un premier séminaire précurseur sur les `Big Data`, le paysage s'est transformé. Aristote a voulu ainsi poser une autre pierre sur l'analyse du phénomène, de ses dérivés, de ses succès. Ce séminaire aura une connotation plus technique. Entre `Hadoop` qui vient immédiatement à l'esprit, les questions de stockage qui connaissent quelques révolutions — disques remplis à l'hélium pour limiter la consommation électrique et la production de chaleur, et d'autres qui vous serons dévoilées par les constructeurs — il reste encore beaucoup de sujets à traiter sur la façon de gérer ces données sur un campus et parfois aussi entre sites. C'est alors qu'apparaît une question majeure : *quid* du transport ?

1.2 Programme

09h00-09h25	*Accueil des participants, café*	
09h25-09h30	**Jean-Michel Batto** (INRA) **Pierre Léonard** (INRA)	Présentation du séminaire
09h30-10h10	**Vincent Heuschling** (Affini-Tech)	L'écosystème d'Hadoop et de ses concurrents
10h10-11h00	**Patrick Demichel** (HP)	Évolutions technologiques pour le Big Data
11h00-11h20	*Pause café*	
11h20-11h50	**Guy Chesnot** (SGI France)	In memory
11h50-12h30	**Patrick Fuhrmann** (Desy)	dCache un système de gestion de données réparties
12h30-13h00	**Jérémie Bourdoncle** (NoRack)	Un système de stockage capacitif *green* et accessible
13h00-14h00	*Déjeuner (salon de marbre)*	
14h00-14h30	**Bastien Legras** (Google)	BigQuery, le Big Data par Google
14h30-15h00	**Patrick Marques** (HP)	Un cas concret chez nos clients : Hadoop
15h00-15h30	*Pause*	
15h30-16h00	**Peter Livaudais** (ParStream)	ParStream, une base de données qui révolutionne la recherche en masse
16h00-16h30	**Sébastien Noury** (Palantir)	Palantir Gotham, une plate-forme d'analyse issue de la Silicon Valley
16h30-17h00	**Philippe Martin** (Dell)	Peut-on faire passer des Big Data avec un modem 56kb/s
17h00	Discussion avec les orateurs, fin du séminaire	

Chapitre 2

Compte-rendu de la journée

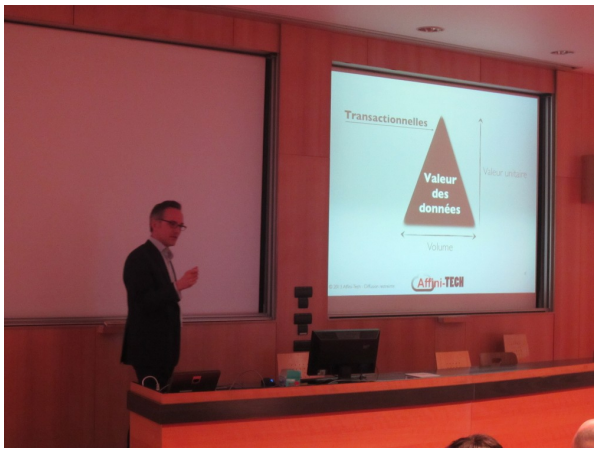
Ce compte-rendu — texte et photographies — a été réalisé par Fabien Nicolas de l'agence Umaps, « Communication de la recherche et de l'innovation », http://www.umaps.fr.

À la poursuite des Big Data

Jean-Michel Batto présente ce séminaire en rappelant qu'Aristote est depuis 1988 une association loi 1901. Grâce au mécanisme d'autogestion, ils ont pu organiser des séminaires sur des sujets précurseurs, notamment sur les Big Data il y deux ans de cela. Aujourd'hui, ils réitèrent l'expérience et font revenir sur l'estrade des intervenants du premier séminaire, ainsi que nombreux autres acteurs qui ont investi le domaine du Big Data.

L'écosystème d'Hadoop et de ses concurrents

Vincent Heuschling (Affini-Tech)



Vincent Heuschling ouvre le séminaire en soulignant combien il est rare d'entendre parler de Big Data sans qu'Hadoop ne soit mentionné. Sa société du nom d' Affini-Tech est spécialisée dans l'infrastructure informatique et Hadoop est au cœur de leur métier, depuis la collecte des données jusqu'à leur présentation, en passant surtout par leur valorisation. Conscient d'introduire une vaste journée sur ce thème, Vincent Heuschling propose donc de se demander avant tout quelles performances sont possibles avec Hadoop, quelles évolutions et surtout, quels usages sont intéressants lorsque l'on n'est pas un colosse comme Google ou un Amazon ?

Mais qu'est-ce que le Big Data ? Des clients d'Affini-Tech croient souvent qu'il s'agit de réunir en un seul endroit leurs données éparpillées. Donc de faire comme avant, mais à plus grande échelle. Or le Big Data, c'est autre chose. Ce

sont des projets pour créer de nouvelles opportunités de business grâce à l'agrégation de données internes et externes à l'entreprise. Le Big Data est une source d'innovation.

Lorsque les données en question sont structurées, issues de base de données classiques, les méthodes habituelles peuvent maitriser leur croissance. Mais il y a aussi les données non structurées, générées par des machines, qui explosent depuis les années 90, depuis la naissance du *Web*. Désormais, l'informatique moderne crée chaque année davantage de données qu'elle n'en avait produit auparavant dans toute son histoire cumulée.

Les caractéristiques des Big Data se résument en quatre « **V** » et autant de problèmes techniques qu'Hadoop est capable d'adresser. Le plus évident d'entre eux est le **V**olume de données. Mais leur **V**ariété importe aussi, car elle nécessite des outils dédiés à chaque type de donnée, là où autrefois un outil pouvait tout gérer. Ensuite vient la **V**élocité : souvent, dans notre monde toujours changeant, des traitements difficiles de données ne sont utiles que s'ils se font en temps réel. Enfin, la **V**ariabilité des données exige des outils très souples. Nous ne savons pas, demain, quels formats nous aurons à traiter. Mais il y a également un volet économique. Dans une entreprise, un petit nombre de données ont très une forte valeur unitaire. Par exemple toute commande égarée est une perte irrémédiable pour l'activité de votre entreprise. Ces données-là méritent donc de gros moyens matériels pour leur traitement.

Mais en dessous, certaines données sont moins importantes et demandent un investissement plus modéré. À l'extrémité de ce classement, il y a les Big Data, des données dont la valeur unitaire est nulle. Chaque page indexée par Google ne vaut strictement rien ! Toutefois la masse de ces pages fait la force de toute la firme. Ce qui tombe bien, car il suffit pour les gérer de *commodity hardware*, le matériel plus simple possible ! Ici les données elles-mêmes sont remplaçables et c'est le logiciel qui crée de la valeur.

Ce logiciel, il faut alors le choisir. La technologieSQL permet de bonnes performances, mais elles s'effondrent dès lors que l'on augmente le volume des données. Des systèmes massivement parallèles repoussent cette limite, mais leur performance s'effondre forcément à terme. La technologie Hadoop est encore moins performante... du moins sur de faibles volumes de données. Car la force d'Hadoop réside dans sa possibilité de passage à l'échelle de façon linéaire : traiter plusieurs pétaoctets demande, proportionnellement, la même puissance que plusieurs gigaoctets ! Il devient alors enfin possible de manipuler de grands volumes pour des temps de calculs raisonnables.

Comment cela marche-t-il ? Hadoop a été créé par une communauté *Open Source* à partir d'une publication de Google. Le principe est simple : envoyer le traitement vers la donnée plutôt que la donnée vers le traitement. Le logiciel distribue l'information dans un réseau unique, constitué de matériel très simple. Les données sont réparties selon le HDFS, *Hadoop File System*, et lorsqu'un traitement est requis, le logiciel envoie ce traitement à chaque machine, qui l'applique aux données qu'il possède. Les résultats du calcul de chaque bloc, légers, sont collectés par le logiciel central. Dans une seconde phase, celui-ci va synthétiser tous ces intermédiaires pour obtenir, enfin, le résultat final. Ce processus en deux étapes se nomme le *MapReduce*. Au cœur d'Hadoop, cette méthode permet de pouvoir tester un traitement sur un ordinateur portable et, le lendemain, de déployer mille nœuds avec le même traitement.

Cette grande puissance vient avec des contraintes importantes. Différentes variantes sont donc développées ces temps-ci, pour gérer efficacement chaque type de données. Ainsi Flume ou Sqoop, par exemple, et chacun peut construire sa propre solution, voir mixer avec des choses comme Cloudera, Impala, Apache Tez (qui évite les I/O sur le disque), Spark (sur un principe innovant d'itérations !) Mais on garde à chaque fois la compatibilité Hadoop.

Hadoop est un concurrent des *Data Warehouse*, mais il s'y intègre aussi progressivement et vise même à les englober. Hadoop est déjà l'outil d'ETA (extraction-traitement-affichage) pour les données des *Data Warehouse*. Au final, les *Datamarts* pourraient être inclus dans Hadoop.

À l'avenir, il est prévisible que la charge de travail des systèmes Hadoop sera augmentée. Un

processus en arrière-plan de la machine traitera quelques pétaoctets par jour. Et d'autres systèmes géreront le flux en continu pour des interactions instantanées avec les utilisateurs.

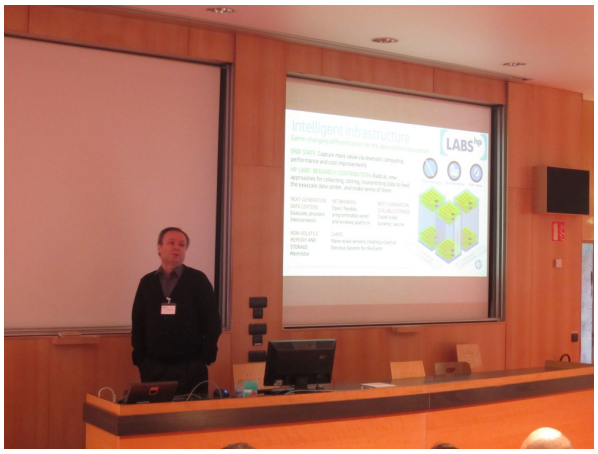
Pour la puissance, on peut d'abord mixer du Hadoop (traitant un milliard de données) et leSQL comme frontal d'accès aux données (gérant le millier obtenu via Hadoop). Aujourd'hui, Facebook est dans un cluster Hadoop de 100 téraoctets de données, avec 100 000 tâches qui tournent en permanence. Mais tous leurs analystes sont devant des outils SQL pour analyser les données.

Voyages SNCF réparaient les incidents de ventes et donnaient autrefois un nouveau billet au client. Ils ne pouvaient pas retrouver l'ancien à travers les 13 serveurs où les commandes passent ! Aujourd'hui, une traitement lancé toutes les 5 min synthétise les actes de ventes et l'on peut revenir dessus.

Les applications de cette technologie sont diverses, allant même jusqu'à l'apprentissage (*machine learning*). Même en France, Hadoop commence à prendre. Si en 2012 la curiosité était de mise, aujourd'hui même les DSI veulent être force de proposition. Hadoop n'est plus seulement une opportunité, cela va devenir la règle.

Évolutions technologiques pour le Big Data

Patrick Demichel (HP)



Chez Hewlett-Packard, la vision du futur est prête jusqu'en 2015. En décrivant les idées de son entreprise, Patrick Demichel, *strategic systems architect* chez HP, propose donc un véritable voyage vers le futur.

Pour HP, l'explosion des données liées au *Web* signifie devoir fournir encore plus de machines pour soutenir la croissance. Sans compter les nouveaux capteurs qui à l'avenir augmenteront encore le volume des données. Le problème est que le support matériel de ces données n'est pas adapté. Aujourd'hui, il faut une journée entière pour lire un disque de 3 téraoctets. Quant à mettre en mémoire d'aussi grands jeux de données pour y accéder plus vite, la technologie *flash* ne permettra pas d'augmenter beaucoup plus les volumes.

Que faire ? Un prix faible par gigaoctet sera indispensable, étant donné les volumes concernés. Et il faut se donner la capacité de trouver où est la valeur dans ces montagnes de données. Deux piliers technologiques seront indispensables : le photonique pour le transport et la mémoire non volatile pour le stockage.

Tous les fournisseurs travaillent actuellement sur ces deux solutions. Dans dix ans, les architectures informatiques seront méconnaissables. Les ruptures technologiques seront profondes, et heureusement ! Car les théoriciens des algorithmes pensent que les algorithmes du futur auront encore moins de localité qu'aujourd'hui. Le plus gros défi n'est pas le contenu, mais les données. Il faut un « *Data-Centric Data Center* » !

Le plan d'HP est en deux phases, dont la première est imminente.

Le transport photonique Tout d'abord, le photonique. Le cuivre ne donne pas assez de bande passante et de puissance. D'ici une décennie, faire face à l'accroissement des données deman-

dera de multiplier par un facteur 30 le débit de données tout en divisant par 10 la consommation électrique. Cela revient à dire qu'en 2020, il faudra consommer 1 picojoule par bit.

Les technologies avec de telles performances fonctionnent actuellement dans les laboratoires... Mais à quel prix ! Le cuivre défie pour l'instant toute concurrence à ce niveau et pour le remplacer, il faudra une technologie encore moins chère.

Un des avantages du photonique est la consommation d'énergie. Le transport n'occasionne aucune perte : peu importe la distance, il n'y a plus de répéteurs comme sur du cuivre. La conversion électrique-photonique est également peu gourmande. HP va perfectionner ses technologies, réduire les coûts et selon leurs plannings, dans un an ils disposeront d'un *backplane* optique. Il permettra de passer toute la connectique en optique sur les processeurs. Ensuite, la dernière étape sera de passer en optique le processeur tout entier.

Ils vont créer l'HyperX Network, qui aura la capacité d'atteindre l'exaflops. Les nœuds y sont organisés en hypercubes, et il y a toujours un câble entre deux nœuds dans une même rangée. La latence est courte. Si l'on s'organisait en tore, ce serait long et coûteux en énergie. Dans l'HyperX, chaque nœud peut communiquer avec les autres de plusieurs façons différentes. Quand un câble est saturé, le transfert passe juste par un chemin un peu plus long, un nœud supplémentaire. Pour atteindre de bonnes performances, il faut beaucoup de nœuds. Il faut donc de très nombreux ports, au moins 128. Cette topologie très optimisée, grâce à la technologie du nœud optique, est économe en liens optiques.

HP pense vraiment être sur la bonne piste. Leur plan de nœud à l'échelle exa n'a pas bougé depuis 2008, une éternité en informatique ! Le photonique est ce qui va effectivement multiplier par 30 la bande passante et diviser par 10 la consommation.

La mémoire non volatile L'autre facette du progrès est la mémoire non volatile. Grâce à elle, les programmeurs ne seront enfin plus au pied du au mur pour rester dans de tout petits caches. Leur premier espoir est le *memristor*, « memory resistor » . Trois compétiteurs sur le marché travaillent déjà dessus, il ne fait aucun doute que nous allons très vite le voir arriver. Il promet des coûts bas et une rétention d'information sur près de dix ans.

Le cycle d'écriture sera un point plus difficile pour les mémoires non volatiles. Les mémoires *flash* supportent 10^5 écritures. Il nous faudra atteindre le 10^{10}. De plus, contrairement aux mémoires actuelles, les futures mémoires seront adressables au bit près. La vitesse n'en est pas améliorée, mais la consommation d'énergie chute enfin, car huit fois moins d'information est lue pour lire 1 bit au lieu de lire 1 octet. Autre caractéristique des mémoires non-volatile : leur latence à peu près égale à celle des mémoires actuelles, d'environ 70 nanosecondes.

Mais ces mémoires non-volatiles fonctionnent à froid ! Tout est simplifié en termes de refroidissement. Évidemment il faudra adapter les *Data Center*, ce qui représente un vaste chantier.

La valeur totale des données du Big Data implique d'utiliser un stockage résistant. De nos jours, on a vu que si une tornade prive d'électricité une semaine un *Data Center*, les données sont détruites. Et les sociétés qui y stockaient leurs données ne valent pas mieux.

Pour le monde des super-ordinateurs, il s'agit aussi de faire des systèmes résilients. La finesse des processeurs rend inévitables les défauts matériels et statistiquement sur un ensemble d'un million, un nœud au moins connaitra un défaut chaque jour. Or avec les nouvelles mémoires, il sera possible de sauvegarder toutes les 5 secondes l'état de la machine tout entière ! Si un nœud est hors service, il continue automatiquement de stocker les messages des autres. Une fois rechargé et réinitialisé dans son état précédent, il les rejouera en s'*overclockant* pour rattraper son retard ! Il suffit de 5 secondes et la vitesse de traitement moyenne n'en est même pas affectée. Pour se prémunir contre les crashs plus profonds, tous les 100 cycles on sauvegardera également sur le nœud voisin, sur le même principe. Et un point crucial de tous ces progrès sera leur quasi-transparence pour le développeur.

Combiner les mémoires non-volatiles et la photonique créera une opportunité totalement en rupture avec le paysage actuel. De nos jours un autre souci réside dans le partage de la puissance de calculs. Si votre nombre de visiteurs et de clients se démultiplie en une heure, il faut aujourd'hui avoir anticipé et loué suffisamment de puissance de calcul. La tranche de puissance payée est donc presque constamment sous-exploitée, gaspillée. Le futur réside dans des travaux facturés exactement pour les ressources effectivement consommées.

`Hadoop`, est une méthode qui permet d'amener les traitements aux données, pour ne plus bouger celle-ci. Car même avec l'optique, déplacer ces masses sera impossible. Mais `Hadoop` n'est pas la seule solution pour cela.

À une échelle globale, le futur réside dans la réplication des données. Nous travaillons notamment sur les *erasure codes* (par exemple le code de *reed-solomon*), où chaque fichier est divisé en 12 parts, mais où 4 parts prises au hasard suffisent pour reconstruire le fichier ! Il y ainsi aussi trois copies de chaque donnée, ce qui permet plus de résilience, mais trois fois plus de bande passante. Ce type de code n'est pas forcément nouveau. Mais de nos jours, les reconstructions passent plutôt par du `Raid 6` (forme d'*erasure code* en mode bloc) et il faut des jours et des jours pour réassembler les données. Notre objectif est de reconstruire en une seconde un disque perdu, avec en plus une priorisation possible des fichiers par rapport à un mode bloc.

Dans les autres améliorations à venir se trouvent les réseaux de neurones, les nouveaux processeurs ou encore les senseurs et caméras à contraste automatique. Mais toutes ces technologies vont chacune générer des volumes gigantesques de données, qui nous ramènent encore au `Big Data`. C'est une course à la performance.

In memory, Guy Chesnot (SGI France)

D'après Guy Chesnot, architecte chez `SGI` France, il y a trois difficultés dans la gestion de données. La première est l'acquisition de ces données, celle dite phase d'ingestion ; la deuxième étape est le stockage des données, toutefois optionnel. L'essentiel reste la troisième étape, l'analyse.

L'ingestion permet de structurer les données. Souvent, il s'agit de prendre des données préexistantes et donc d'adapter leur structure. Mais quand les données sont neuves, il faut non seulement les organiser, mais aussi veiller à leur qualité. Les données erronées peuvent poser de gros problèmes.

Pour l'analyse des données, le volume est problématique. Les entreprises ont dépassé le gigaoctet et le téraoctet et il n'est plus rare de voir des pétaoctets (10^{12}) de données. Sans parler de l'exaoctet (10^{15}), qui n'est plus si loin. Pour gérer des transactions commerciales en ligne, il faut pouvoir manipuler de telles masses de données.

Comment réussir de telles analyses ? Les besoins se classent en deux grands axes. Le premier est la connaissance de la question : par exemple, dans le cas d'une recherche de corrélation entre des données, il est impossible de savoir à l'avance ce qui est à trouver. Le deuxième besoin possible est celui de l'aiguille dans la meule de foin. La valeur recherchée est connue, mais il faut découvrir à quoi elle correspond. En fonction du besoin, il faut choisir une architecture adaptée.

Analyse : le cas des questions connues Les architectures à mémoire distribuée, comme Hadoop, ne se maitrisent pas facilement. Elles peuvent surtout traiter surtout les problèmes rentrant dans la première catégorie, dont la question est déjà connue. Car un environnement Hadoop nécessite environ 600 paramètres à configurer en fonction de l'objectif, dont au moins 100 sont indispensables ! Il faut savoir exactement ce que l'on recherche.

Un bon exemple est donné par Netflix, un service américain de *streaming* vidéo à la demande sur une sorte de décodeur TV. Les données issues des décodeurs TV sont collectées, mais pas stockées, et le service est à même de formuler des prédictions en temps réel, comme une invitation à regarder la saison 3 de la série *The Big Bang Theory* quand vous avez fini la saison 2, ou encore de suggérer un abonnement plus adapté à vos goûts, en fonction de ce que vous avez regardé. Mais il s'agit aussi d'une chaîne complexe de distribution de contenu. Contourner les besoins d'analyse, quand cela est possible, ne résout pas le problème à cause de l'accumulation des traitements. Une analyse temps réel est nécessaire et la méthode *MapReduce* d'Hadoop le permet, grâce à l'alternance des phases.

Analyse : le cas des questions libres Ensuite vient le cas où, la question est inconnue. Guy Chesnot cite l'exemple d'une *start-up* qui travaille sur la sécurité des avions au décollage : l'objectif est de savoir ce qu'il faut faire et... ne pas faire pour minimiser le danger de ces manœuvres. Il faut alors analyser la totalité des vols, sauf qu'il n'y a en général pas de paramètres anormaux dans les mesures. L'élément recherché, la cause des dangers, est un signal faible en termes de théorie du signal. La cause des dangers réside-t-elle dans la pente, le vent, la courbe de virage ? La réponse n'est pas connue à l'avance et ne peut donc émerger que d'un très grand nombre de données.

C'est d'ailleurs ce besoin de données massives qui rend essentiel de tout conserver et cela n'est pas forcément bien compris par certains industriels.

Il existe une confusion entre le *Data Mining* et le Big Data. Le premier consiste à excaver une information d'un ensemble de données, tandis que l'enjeu du second est de traiter cinquante mille ensembles de données sans *a priori*. Cela n'a rien à voir.

Guy Chesnot illustre la finalité des Big Data avec l'exemple des jeux olympiques : tous les deux ans, les médias de chaque pays commentent les jeux, mais ils ne parlent quasiment que de leurs sportifs nationaux. Comment avoir une vision globale des jeux ? Une analyse classique des entrées et des sorties des bases de données de chaque pays est impossible : on ne cherche pas une aiguille sur 50 meules de foin. Il faut nécessairement cumuler les informations de toutes les sources et c'est tout l'intérêt de l'analyse à mémoire partagée.

Si les questions connues correspondent à la recherche d'une aiguille dans une meule de foin, les questions inconnues seraient plutôt une comparaison entre les brins de paille provenant de multiples meules.

Or pour ces questions plus floues, où il est nécessaire de relier de multiples meules de données, Hadoop n'est pas adapté. Des solutions à mémoire partagée, comme SGI UV, tirent bien mieux leur épingle du jeu.

Après l'analyse, la visualisation Toutefois, même après l'analyse, la réponse obtenue n'est pas toujours binaire, très tranchée ni même compréhensible telle quelle par des êtres humains. L'idéal serait d'arriver, après la visualisation, à une décision binaire : oui ou non. En Italie, PayPal en fournit un bon exemple avec ses outils de détection de fraude. Ils compilent un très grand nombre de transactions pour décider si la transaction suivante est correcte ou incorrecte.

Mais le plus souvent, les résultats des Big Data doivent être visualisés. Le volume des données oblige à avoir une vision globale, qui passe par des graphiques, des cartes ou des schémas. Certaines visualisations se font avec des cartes de chaleurs (*heatmap*), comme celles à base de

tweets. En mémoire disque, une telle tâche nécessiterait de gérer 300 000 *tweets* par minutes... un vrai défi pour le futur.

Enfin, certains se demandent comment garantir la cohérence des résultats d'un algorithme comme le *MapReduce*. Le parallélisme peut générer des absurdités s'il est mal maîtrisé. Mais Guy Chesnot révèle qu'avec les Big Data, certains outils du monde des statistiques qui étaient réputés fiables ont montré leurs points faibles. Il a été découvert qu'ils divergeaient, même si cela n'était pas clair sur des volumes de données plus faibles... Donc rien n'est certain, les outils anciens ne sont pas plus garantis que les nouveaux dès lors que la volumétrie explose.

dCache un système de gestion de données réparties

Patrick Fuhrmann (Desy)

dCache est un système de fichiers et une aventure qui a commencé en 2000, suite à un effort commun du Fermilab et du *Deutsches Elektronen Synchrotron* (Desy), et qui est aujourd'hui *Open Source*. Patrick Fuhrmann (chef de projet de dCache) cite leurs plus gros clients, comme le European XFEL, qui produit 50 pétaoctets par semaine. Le plus gros de tous est le *World LHC Computing Grid*, avec 100 pétaoctets stockés dans le monde entier. Patrick Fuhrmann souligne combien il est impossible de traiter un volume pareil de données sans un système de fichier adapté. Ainsi le projet LOFAR visait à enregistrer tout ce qui se passe dans le ciel... mais son ambition a dû être restreinte, car cela représenterait vraiment trop de données.

La conception des systèmes informatiques doit être repensée pour dépasser les échelles, pour que les services deviennent indépendants des lieux où les traitements sont effectués. Notre *design* consiste en un mini-processeur dans chaque nœud, puis en une couche intermédiaire pour répartir les données dans les disques. Les services communiquent par un système d'échange de messages. Il devient alors possible de gérer des systèmes de fichiers très réduits, très puissants, ou même très dispersés géographiquement.

dCache effectue plusieurs tâches, à commencer par une séparation de l'espace des noms (*namespace*) et du stockage physique. C'est-à-dire qu'un gestionaire de localisation (*location manager*) distribue les données pour que la localisation physique devienne invisible pour l'utilisateur. Mais le projet détecte aussi les points chauds du réseau, surchargés en données et aux accès fréquents, il déplace alors des données pour calmer la situation. Par exemple, les données sont réparties sur le matériel en fonction de leur type : les vidéos à *streamer* sont sur des disques durs, tandis que les données qui requièrent une écriture et lecture rapide vont sur des SSD. La migration des données rend possible de décharger très facilement des machines pour effectuer leur maintenance ou même un remplacement. Enfin les données sont triées par ordre de valeur. En effet, les DVD peuvent être *re-rippés*, mais les photos ne peuvent pas être reprises ! Les données les plus précieuses migrent donc physiquement vers des supports sécurisés.

Un autre de leurs objectifs est de favoriser les collaborations, comme celles qu'ils entretiennent avec l'EMI (*European Middleware Initiative*) ou Globus Online. En passant, Patrick Fuhrmann re-

commande à tous le service de Globus online, qui permet de déplacer gratuitement des masses de données.

Un des objectifs de dCache est promouvoir les standards. Trop souvent, les grandes compagnies font les choses par elles-mêmes. Elles créent leurs propres formats et il est ensuite très difficile de faire tourner leurs logiciels sur d'autres machines ou de les interfacer. Unifier les protocoles est nécessaire et c'est notamment ce qui nous ouvrira le *Cloud* de demain. dCache supporte GLUE 2, SRM ou encore WebDAV ou StAR. dCache est compatible POSIX.

Patrick Fuhrmann tient également à attirer l'attention sur un autre problème : la puissance de l'*Open Source* a donné aux sociétés la mauvaise habitude de ne plus payer pour le logiciel. Ce sont pourtant ces mêmes logiciels qui manipulent leurs précieux pétaoctets de données et ils le savent puisqu'ils dépensent une fortune pour les serveurs. Cela aboutit à une situation où des initiatives comme dCache sont difficiles à financer.

Il conclut en ouvrant sur les solutions de gestion des identités sur Internet, dont personne ne sait comment elle pourrait être réalisée. Les solutions valables en Allemagne ne seraient pas compatibles avec celle valable en France, or un scientifique a de l'expérience dans plusieurs pays, se déplace de l'un à l'autre. Il appartient à des groupes, mais possède ses publications en propre. FB Connect serait une excellente solution technologique, mais impossible en Europe à cause de la législation européenne, exigeante.

BigQuery, le Big Data par Google

Bastien Legras (Google)

Google a une forte implémentation française, avec 475 employés rue de Londres, dans le 9^{e} arrondissement. Ce premier *GooglePlex* hors zone anglo-saxonne réalise à la fois de la R&D, de la vente, et constitue le premier institut culturel Google en Europe, qui travaille sur numérisation des œuvres.

Google est le 3^{e} assembleur de serveurs au monde, et cela uniquement pour ses besoins internes. Ils soudent eux-mêmes leurs cartes mères, en fonction de leurs besoins. Il y a un peu plus de communication sur ces fameux *Data Center* maintenant, alors qu'avant le secret était jalousement gardé.

Bastien Legras, ingénieur en solution *Cloud*, explique qu'aux yeux de Google, le problème du *Cloud* est de savoir où se trouvent les ressources. *Google Cloud Plateforms* est leur solution. Ils ont mis 4 ans pour la développer, avec notamment l'App Engine, le Compute Engine, le Cloud Storage, Big Query et CloudSQL. L'App Engine fournit une interface pour rendre les choses simples, le Cloud Storage permet un stockage intelligent, basé sur celui que s'était construit Google en interne pour son propre usage, et BigQuery fournit des requêtes ultraperformantes.

La philosophie Google est de rendre les choses simples pour permettre à chacun de se concentrer sur le cœur de ce qu'il fait. Via les API, par exemple sur Facebook, sur Météo France ou via des *tweets*, ils collectent des données. Après traitement, celles-ci peuvent par exemple détecter une

inondation avant toute annonce officielle. L'objectif est alors d'anticiper, par exemple d'adapter tout de suite le *Back Office* et le *Front Office* du service qui aide les victimes pour gérer tous les appels.

Astro Teller, ingénieur chez Google, a formulé à voix haute un des brins de l'ADN de Google : « Il est plus facile de multiplier par 10 que d'augmenter de 10% » . Ce dicton de Google signifie qu'il est plus facile de progresser lorsque l'on fait table rase du passé et que tout est repensé depuis la base. Il s'agit d'une composante importante de leur culture R&D.

Google utilise des Big Data depuis 15 ans et il y a toujours plus de cas concrets, comme le besoin de haute performance, les grands volumes de données ou encore le besoin de réponses ultrarapides.

L'historique d'architecture distribuée de Google commence en 2004, quand Hadoop naîtà partir des publications par Google de GFS et de *Map Reduce*. Celles-ci se trouvent encore sur Internet. Pour Bastien Legras, souvent, soit les produits informatiques naissent des publications Google, soient ce sont des externalisations des produits Google.

En 2006, le Big Table de Google a répandu l'utilisation de NoSQL. En 2010, c'est au tour de Dremel. Si Hadoop est un « 4x4 », très polyvalent, Dremel est par contre bien plus rapide. Et à présent Big Query serait sa suite logique. En 2012, Colossus a été notre impulsion de stockage *Cloud*.

Le *MapReduce* inventé par Google donne accès à des coûts réduits pour effectuer des calculs en réseaux. Toutefois il faut aussi réaliser que *MapReduce* a 14 ans à présent. De nombreux algorithmes voient le jour et beaucoup de progrès ont été réalisés, et les technologies *MapReduce* actuelles réalisent avec dix fois moins de cœurs des travaux plus rapides. Pour beaucoup de tâches, Hadoop et Big Query sont complémentaires.

Le public du séminaire s'inquiète de la durée de vie de Google. L'entreprise ayant été créée il y a seulement 15 ans, comment peut-on être certain de sa fiabilité sur le long terme ? Mais Google se défend de fournir Hadoop ou tout autre service de ce type : ils créent plutôt des briques logicielles pour que chacun déploie ses propres solutions. De nos jours, des applications passent en une semaine d'une centaine à plusieurs millions d'utilisateurs. Jamais un tel développement n'aurait été possible si chacun avait dû posséder son matériel. L'architecture distribuée était obligatoire.

Par exemple, dans le vrai cas de figure d'un site d'*e-commerce*, l'objectif est d'être capable d'afficher en temps réel des propositions en fonction de leur connaissance de l'utilisateur. Un *tracker* tourne donc en permanence, crée un *cookie*, et l'utilisateur est « profilé » en 3 clics ! Selon qu'il a regardé des écrans ou des sèche-cheveux, il faut lui proposer un produit adapté pour qu'il reste sur le site. Bastien Legras souligne même qu'il y a des sociétés entières dédiées à aider à convertir l'utilisateur en client, en choisissant quel produit lui afficher ! Derrière une simple recommandation se tient une société tout entière.

De plus les données se cumulent : 20 000 utilisateurs analysés peuvent ensuite être compilés et visualisés par le *marketing*, pour comprendre le marché.

Ces traitements s'appliquent à tous les domaines. Par exemple des usages pétroliers par forcément attendus : les nappes évidentes ayant été trouvées, des échographies complexes du sol sont requises pour en découvrir de nouvelles. Même 3000 machines ne pourraient pas analyser la masse de données de ces échographies. Pour y parvenir, la méthode développée est un processus itératif : on avance un peu le calcul, puis on regarde les résultats et l'on fait des simplifications à la main ! Après 6 itérations de ce régime, on peut enfin achever le calcul.

Google Compute Engine est développé depuis 3 ans avec 1000 ingénieurs. Microsoft et Amazon proposent des solutions très proches et au vu les investissements réalisés, Google pense faire au moins aussi bien que les autres. Le public s'inquiète toutefois de la domiciliation de Google aux États-Unis... où le *Patriot Act* permet au gouvernement de regarder toutes leurs données ! Les envoyés de Google se veulent rassurants : ces points sont discutés jusqu'à trouver un terrain d'entente adapté à chacun. Il faut de plus noter que toutes les données sont cryptées et que seul le client

aura la clef primaire de la donnée. Donc Google ne peut pas lire les données. Enfin les entreprises se posent moins la question de nos jours, car la plupart ont une filiale aux États-Unis, ce qui de fait, leur donne des obligations identiques à celles de Google vis-à-vis du *Patriot Act*.

Un système de stockage capacitif *green* et accessible

Jérémie Bourdoncle (NoRack)

Deux ans auparavant, *NoRack* était venue parler de ses idées au premier séminaire Aristote Big Data. Aujourd'hui, elles se sont concrétisées par une *start-up*.

À l'origine, l'intervenant Jérémie Bourdoncle — PDG de NoRack — et ses collègues se demandaient surtout comment ils pourraient avoir du matériel pour faire un entrepôt de données sans en louer ni en acheter. Leur idée fut, comme Google le fait, de construire eux-mêmes leur matériel. Ils se heurtèrent alors à la principale problématique des centres de données : le refroidissement pour l'utilisation à temps plein des machines.

Le principe mécanique des serveurs n'a pas changé depuis 30 ans, les *racks* sont les mêmes. Lorsque Jérémie Bourdoncle et ses collègues se sont fait livrer un *rack* pour leur projet, celui-ci pesait 280 kg ! Entièrement en acier, les *racks* semblent conçus comme pour résister à l'impact d'une voiture. Les *Data Centers* eux-mêmes se sont améliorés, avec des allées froides, des allées chaudes, et d'autres choses, mais pas les *racks*, dont la forme standardisée emprisonne la chaleur. Pour 1 watt consommé pour du calcul, aujourd'hui environ 1.3 watt sont dépensés en refroidissement.

NoRack s'est donné pour objectif de repenser le *rack*. Leur première initiative a été d'intégrer les allées chaudes et des allées froides à l'intérieur de la structure. Ensuite, puisque l'on va vers des nœuds toujours plus nombreux, une alimentation générale a été ajoutée, multiple et optimisée. Le châssis mécanique de ce « non-*rack* », breveté, réunit à chaque étage huit machines disposées en octogone autour de la cheminée. L'air ambiant froid est aspiré, refroidissant les machines et concentrant la chaleur dans la cheminée centrale où il est évacué. Il atterrit alors *via* un tuyau à l'extérieur du bâtiment. Et tout cela n'est animé que par un petit ventilateur, car l'effet Venturi fait le reste. Le besoin de climatisation est totalement supprimé et la consommation électrique en est terriblement réduite. Les alimentations sont à base d'ATX, classiques sur le marché.

L'architecture est en grappe (*cluster*) : chacune des huit plaques de chaque étage a des emplacements pour un processeur et cinq disques. Chaque plaque peut se retirer individuellement, ce qui permet d'accéder aux machines bien plus facilement que dans les *racks* actuels.

Des cas d'usages ont été réalisés, notamment pour des usages de services Internet. Il s'agissait de faire varier la charge des nœuds, pour voir comment gérer les pointes de trafic. Au final, leurs serveurs ont eu 99,2% de disponibilité, sans avoir un énorme *data center* autour d'eux, dans une salle oscillant librement entre 10 et 30 degrés. Si un cœur commence à chauffer, le ventilateur accélère simplement un peu. La consommation n'était que de 5 kW pour 750 téraoctets, soit bien moins que pour une infrastructure traditionnelle. Une performance réalisée avec une machine utilisant 20% de la matière d'un *rack* classique, et donc en réduisant le prix.

Il s'agit aujourd'hui d'adapter les *racks* aux nouveaux usages, y compris le Big Data.

Un cas concret chez nos clients : Hadoop

Patrick Marques (HP)

HP reprend la parole avec Patrick Marques, architecte avant-vente chez HP, pour détailler plus précisément leur utilisation de Hadoop. Plate-forme ouverte, Hadoop passe à l'échelle (*scalability*), est distribué et résilient. Pour cela, il repose d'un côté sur le HDFS et de l'autre sur *MapReduce*. Cela dit, Hadoop est écrit en Java, un langage pour lequel l'engouement est en baisse. En réalité, à partir des deux bases s'est développé tout un écosystème de produits *Open Source*, tels que Pig, Hive, Hbase, Flume, Oozie, Sqoop et bien d'autres. La force de tous ces produits Hadoop est d'amener le traitement jusqu'à la donnée, donc de transporter seulement quelques kilooctets de code. Cette localité des données leur confère une capacité à monter en charge de façon linéaire : multiplier par 10 les données ne multiplie que par 10 les calculs.

Quels sont ces concepts faisant fonctionner Hadoop ? Le HDFS tout d'abord distingue deux types de serveurs : ceux pour les métadonnées (*NameNode*) et ceux pour les données (*DataNode*). Or, les disques étant organisés en mode Raid, c'est toujours le disque le plus lent qui limite la vitesse d'un transfert de données. Ainsi, dans Hadoop, pour une bonne performance, les métadonnées sont chargées dans la RAM. La taille de celle-ci détermine alors le nombre de fichiers adressables. Chaque donnée est également divisée en blocs (*splitting*), souvent de 64 ou 128 mégaoctets, chacun est par défaut répliqué trois fois, pour plus de performance et de disponibilité. Enfin Patrick Marques estime important de noter qu'il ne s'agit pas d'un système POSIX, et qu'il est basé sur le GFS de Google.

Pour visualiser le fonctionnement du *MapReduce* d'Hadoop, Patrick Marques propose d'imaginer que l'on veuille compter les mots d'un texte. L'étape *Map* fait compter à chaque machine tous les mots de toutes les phrases qu'elle a en mémoire, puis l'étape *Reduce* va sommer tous ces comptes intermédiaires pour obtenir le total. Toutefois il avoue que concrètement le processus est plus compliqué, car il y a une étape intermédiaire, souvent nommée *shuffle & sort*. D'un point de vue applicatif, Hadoop implique un développement dans un langage de haut niveau et une abstraction de l'architecture matériel, dans une API Java.

La transition aujourd'hui est importante, analogue à celle du passage du HPC aux grappes de calcul (*clusters*), dans les années 90. De son côté, HP a déjà construit le CMU, (*Cluster Management Unit*, qui peut provisionner des milliers de serveurs dans le monde, les surveiller (*monitoring*) et les administrer en temps réel.

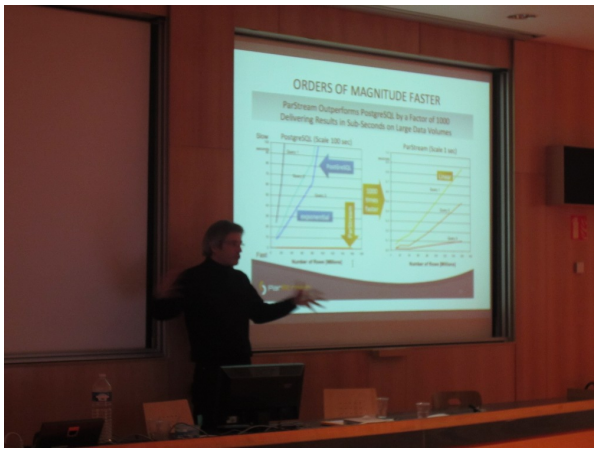
Mais comment dimensionner son réseau ? Quel nombre de disques par *socket* et quel nombre de téraoctets par *rack* ? HP s'est choisi un standard en étudiant ses retours clients : ils ont opté pour des machines Linux monocœur dotées de 4 gigaoctets de mémoire. Ils veillent à rester sur des disques de 3,5 pouces, car ceux-ci ont un meilleur coût. La RAM de 4 gigaoctets est une bonne base et ils se contentent ensuite de varier autour de ce point d'équilibre. À budget équivalent, il faut choisir entre avoir beaucoup de serveurs ou bien de meilleurs serveurs. Si le parc n'est pas bien dimensionné, le risque est l'étranglement à la moindre surcharge de travail. Souvent dans les

entreprises, beaucoup de puissance est déployée juste pour pouvoir amortir ces pics, mais avec Hadoop, il devient possible de dimensionner pour avoir juste ce qu'il faut. Hadoop donne du sens à la donnée, il ne la stocke pas mécaniquement.

Et bien entendu, Hadoop fait des petits, comme les scalable NAS, Moose FS, Gluster FS ou encore Scality et Openstack.

ParStream, une base de données qui révolutionne la recherche en masse

Peter Livaudais (ParStream)



Peter Livaudais, directeur des développements de la société ParStream, rappelle aux auditeurs du séminaire que le Big Data n'est pas une nouveauté. Il y a quarante ans, Herbert Simon, futur prix Nobel, s'étonnaîtdéjà de l'explosion des données... Plus récemment, Jeff Hammerbacher, un des créateurs de Hadoop disait « qu'il était impossible de suivre tout ce qui se passait dans le domaine ». Or c'était il y a deux ans, avant que tout ne commence !

À présent, quelque chose émerge de l'activité autour du Big Data. Daniel Crobb, qui enseigne sur les systèmes complexes à l'École Polytechnique, utilise l'exemple des briques : bien que posées les unes après les autres, elles forment peu à peu quelque chose de plus, un mur. Le Big Data créé quelque chose de plus, quelque chose de sociotechnique, des changements aussi grands que ceux connus avec Internet ou le téléphone mobile. De nouvelles façons de traiter les problèmes.

Peter Livaudais pense que dans ce séminaire le sujet du transport des données au niveau du disque a été bien traité, mais que l'on n'a pas entendu parler des transferts à l'échelle macro, c'est-à-dire les mouvements de données à grande échelle. Ni d'ailleurs à l'échelle micro, sur le processeur et le cache. Notamment, comment satisfaire les besoins de l'analyse en temps réel de Big Data ? La solution ParStream tient en 3 mots clés : **index**, **compressé**, **parallèle**.

Le projet est né en 2007, à l'époque une équipe d'ingénieurs allemands s'est vue demander d'adresser une base de 18 milliards d'enregistrements avec 1000 requêtes par secondes et des temps de réponse de 20 millisecondes (ce qui est tout juste deux fois le temps d'un accès disque...). Il a été nécessaire d'aller au-delà de ce que le marché proposait. Experts en machines mobiles, ils considéraient que l'essentiel n'était pas tant le stockage de la donnée, mais la restitution de la réponse à une requête.

L'index compressé qu'ils ont breveté est à la base de la performance de ParStream. Contrairement aux autres solutions, il n'est pas nécessaire de le décompresser pour pouvoir le lire. La recherche va se faire directement sur du compressé, d'où un gain de vitesse conséquent. Sans compter que chaque requête utilise des cœurs multiples, ce qui permet enfin d'être linéaire, de gérer les taches en parallèle et donc encore plus vite. En termes de performances, PostgreSQL est dépassé d'un facteur 1000 lorsque de grands volumes de données sont traités.

ParStream vise de grands volumes comme Hadoop, mais en temps réel. Une de ses premières applications fut d'ailleurs la base de données du New York Times. Les cas d'usages sont variés. En Australie, ParStream est utilisée pour des sondes minières, après l'extraction, pour assembler

les minerais en catégories. On l'utilise également en recherche climatique ou, dans un tout autre domaine, *via* des médias comme Search Metrics. L'avion, la voiture, tout change et même les outils manuels sont désormais assujetis aux Big Data.

Leur logiciel peut être installé *via* le *Cloud* ou par des partenaires comme OEM ou ISV. Et son prix ne dépend que du volume maximum de données que le client souhaite y placer. Les tests sont gratuits.

Palantir Gotham, une plate-forme d'analyse issue de la Silicon Valley

Sébastien Noury (Palantir)

Comment utiliser le Big Data ? Pour Sébastien Noury ingénieur chez Palantir, c'est une source de données avec lesquelles il faudra régulièrement interagir. Au-delà des automatismes, la clef est de fournir aux humains des outils pour analyser les données en temps réel.

La réalisation la plus médiatique de son entreprise, Palantir, est l'outil qu'ils ont fourni à la Team Rubicon. Formée après l'ouragan Sandy, cette équipe recensait les habitants qui avaient besoin d'aide à cause de la catastrophe. Grâce à Palantir, toutes ces demandes ont pu passer par une application mobile, faisant gagner un temps et une énergie précieux aux secours. Palantir doit d'ailleurs une fière chandelle à l'opérateur téléphonique AT&T, pour avoir déployé des stations roulantes pour maintenir en état de marche le réseau mobile.

Les fondateurs de Palantir sont des anciens de chez Google. Après l'avoir quitté, ils ont travaillé pour l'équipe anti-fraude de PayPal, société acquise par eBay en 2002, lors de l'explosion des transactions en ligne. PayPal, après avoir évalué les limites d'une détection de la fraude robotisée, a alors choisi d'augmenter la part d'intelligence humaine dans son système de détection de fraudes. La principale raison pour cela était qu'aucun système automatisé ne pouvait traiter un tel volume de transactions sécurisées, et de là est née l'idée fondatrice de Palantir, la symbiose entre l'homme et la machine. En 2004, la société Palantir était fondée.

La plate-forme Palantir Gotham vise à aider les analystes à traiter des volumes de données qui, sinon, ne donneraient qu'une vision très partielle de la réalité. L'approche est itérative : le logiciel traite les données, un être humain réfléchit au résultat et relance une analyse plus appropriée, ce cycle se répétant jusqu'à obtenir le résultat recherché.

Palantir a une structure inhabituelle, car elle est constituée d'ingénieurs dans une proportion très importante (plus de 80%). Ce sont eux qui vont sur le terrain et font les démarches habituellement segmentées en différent corps de métier. C'est une approche résolument tournée vers l'ingénierie, pour répondre à des besoins très précis dans un laps de temps le plus court possible.

Il était nécessaire d'abolir les barrières, de défragmenter les sources de données et de démocratiser des outils d'analyse puissants. La solution était que les ingénieurs puissent voir les problèmes pour y répondre de façon efficace, et être en symbiose sur le terrain avec les experts des problèmes concernés, pour les résoudre avec efficacité.

Un de leurs partenariats le plus notable est celui, très actif, avec les forces de police de New York et de Los Angeles, la plus importante police au monde. Leurs missions les amènent à surveiller la prolifération d'épidémies, à mesurer l'impact des catastrophes sur l'environnement, à lutter contre des fraudes et des intrusions informatiques, ou même à démanteler des réseaux de trafic de drogue. Il s'agit souvent de filtrer en une seconde des jeux de données très bruités.

Les outils de Palantir permettent par exemple, à partir de deux adresses IP responsables de connexions suspectes, de cumuler critère d'analyse après critère d'analyse pour reconstruire le système dont elles proviennent, pour visualiser et comprendre l'attaque informatique qui a eu lieu. Leur système est pensé pour que l'analyste conduise la machine dans l'enquête, dans une sorte de symbiose. Ils sont capables de gérer des gigaoctets, mais aussi des pétaoctets, presque en temps réel, avec une actualisation à chaque minute.

Peut-on faire passer des Big Data avec un modem 56kb/s

Philippe Martin (Dell)

Pour Philippe Martin ingénieur réseau avant-vente chez Dell, dans le Big Data, la priorité est trop souvent donnée au nombre de nœuds, pour analyser un maximum de données possibles. Le réseau est souvent l'oublié de ces projets et ses performances pèchent. Depuis plus de dix ans, Dell construit ses propres éléments réseau. D'une activité ciblée et simple, leur part réseau s'est accélérée depuis deux ou trois ans et ils sont désormais le quatrième constructeur mondial de *switches*, toutes catégories confondues.

Leur objectif réseau est de proposer une alternative simple, ouverte, performante et sécurisée. L'ouverture est vitale : chacun doit pouvoir changer de constructeur quand il le souhaite, mais le matériel doit également être ultra-performant et ultra-sécurisé. La vision développée par Dell pour répondre à ces points se nomme Active Fabric.

Cette approche délaisse les produits en mode châssis avec cartes d'extensions, trop coûteux, limités en termes de connectivité et trop demandeurs de maintenance. Dell les remplace par des boitiers pleinement garnis moins coûteux et surtout avec des protocoles d'accès plus ouverts. Ils tablent sur un cœur de réseau distribué, qui coûte seulement un peu de bande passante aux machines, mais augmente vraiment la résilience du réseau, réduit l'impact coût du châssis et améliore l'évolutivité du cœur de réseau.

Quels produits réseau permettent ces Big Data avec cœur de réseau distribué ? Tous les équipements Dell ne se bloquent pas même si on les utilise à pleine capacité, et ils fonctionnent tous sur les technologies actuelles. Philippe Martin présente quelques modèles de la gamme Dell pour les infrastructures de Big Data : le Z9000 est leur gros produit pour le cœur de réseau, avec ses 32x40 Gb, Microsoft en utilise 128 pour faire fonctionner son moteur de recherche Bing. Il présente un ensemble de produits cohérents. Dell a également développé des solutions logicielles pour gagner du temps, notamment Active Fabric Manager, qui va permettre d'aller paramétrer l'ensemble des équipements qui constitue le cœur de réseau, et de déployer rapidement de nou-

veaux équipements sur le réseau. L'outil OMNM de Dell va quand à lui permettre de réaliser une cartographie de l'infrastructure existante, un inventaire très complet, mais automatisé ! Il permet aussi de faire une remontée de statistiques de niveau 3, sFlow, NetFlow, OpenFlow et IPFIX, et de les visualiser.

Que peut-on proposer de plus que des produits plus ou moins performants ? Aujourd'hui le sujet est l'automatisation, pour demain certains réfléchissent à l'allocation dynamique de ressources. Philippe Martin présente quatre perspectives importantes pour Dell. Tout d'abord le *Bare Metal Provisonning*, capacité d'un équipement à peine branché sur le réseau à aller chercher son fichier de configuration, se déployant ainsi très rapidement. Le *Smart Scripting* permet de faire tout tourner sur tel équipement, du moment qu'il s'agit de Perl ou de Python. La virtualisation, demain avec HyperV et après-demain avec KVN, est la capacité des équipements Dell à être dynamiquement configurés par une solution d'hypervision, lors de déplacement de machines physiques. Enfin le *Programmatic Management* consiste aussi en une configuration dynamique, mais contrôlée par une solution d'orchestration.

Aujourd'hui, plus l'infrastructure est importante, plus la multiplicité des boites, serveurs, châssis, dispositifs de stockage et outils rend son utilisation complexe. Les constructeurs comme Dell travaillent donc à ce que le réseau puisse faire abstraction du matériel utilisé. C'est ce qui est appelé l'approche *software defined networking*. OpenFlow en est le représentant le plus connu, qui touchera vraiment les acteurs d'ici deux à cinq ans. L'idée est que l'équipement réseau est constitué de deux parties : d'un côté le *Data Plane* qui avec sa capacité de traitement peut être vu comme les bras de la machine ; et d'un autre côté le *Control Plane* qui est comme le cerveau et contient le protocole d'interaction entre les boites. Manier le mélange des deux requiert une grande expertise : Open flow les sépare donc. C'est une solution extrêmement intéressante. Dell ne vise pas à faire des « demi-*switches* » un peu bêtes, mais travaille sur la capacité de leurs équipements à s'intégrer dans des environnements complexes, de façon extrêmement simple.

Chapitre 3

Présentations

3.1 Jean-Michel Batto, Pierre Léonard (INRA)

Présentation du séminaire

A la poursuite des Big data

Jean-Michel Batto / Pierre Léonard

Aristote est une société savante et philotechnique qui regroupe depuis plus de 25 ans organismes de recherche, grandes écoles, entreprises et PME impliqués dans les nouveaux développements et usages des technologies de l'information :
13 EPST
2 écoles
10 sociétés dont 5 pme

Le CPG Comité de programme et ses groupes de travail est une instance de réflexion et de débat. C'est aussi le lien indispensable entre l'association et les organismes membres mais aussi les auditeurs de nos séminaires

aristote

Séminaire du 27 Mars 2013

Traitement des données

Elastic
M-R

dryad

Infrastructures nécessaires

Le grand oublié

aristote

Séminaire du 27 Mars 2013

Du discours vers les réalisations

Ce jour 27 Mars 2013

aristote

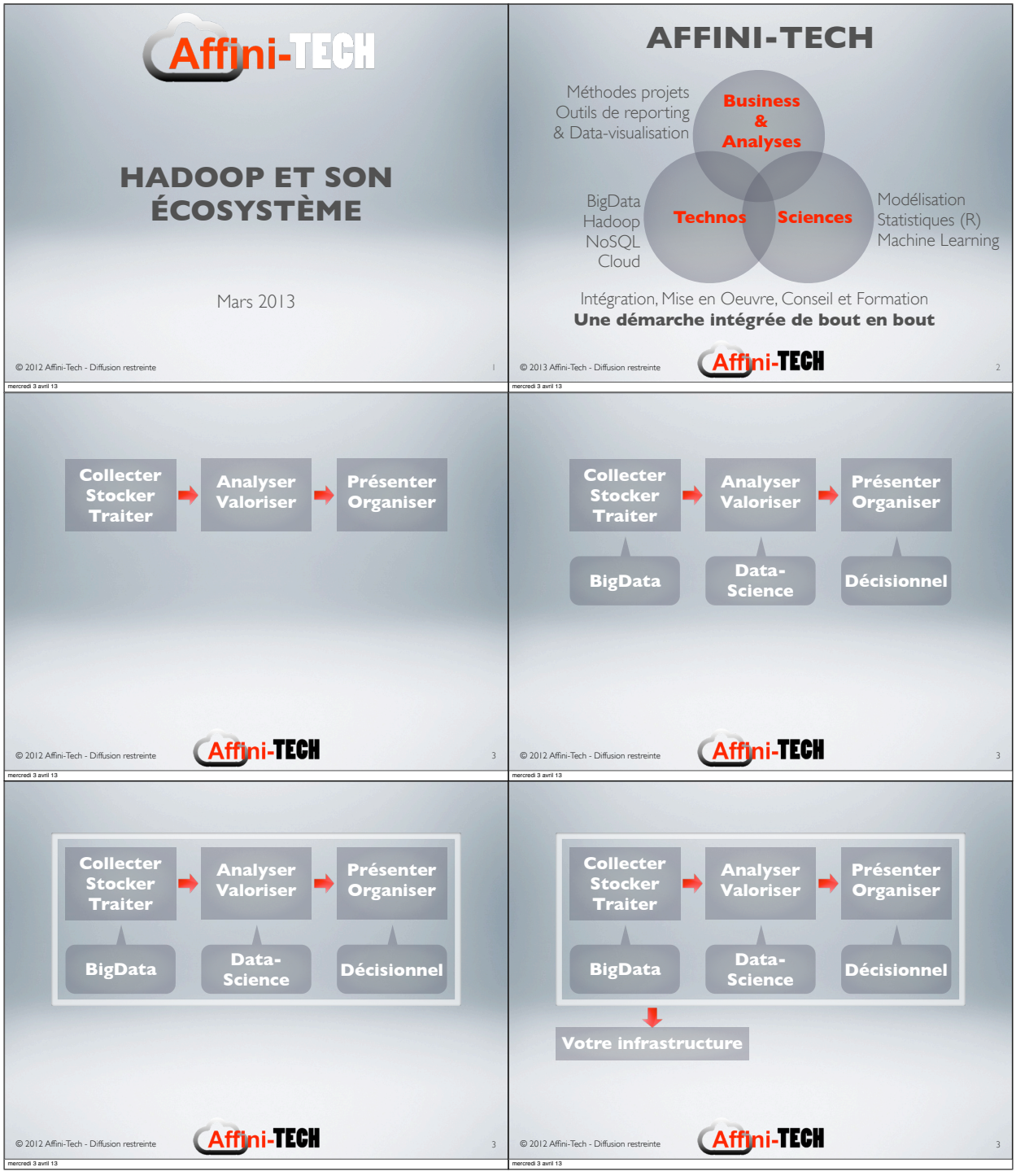
Séminaire du 27 Mars 2013

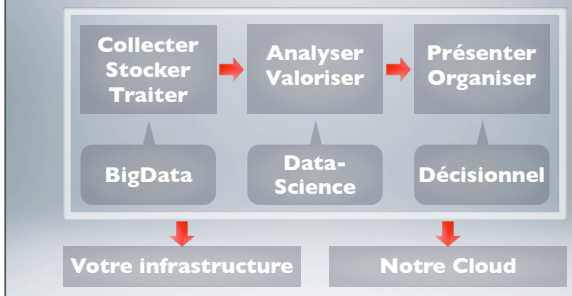
3.2 Vincent Heuschling (Affini-Tech)

L'écosystème d'Hadoop et de ses concurrents

L'écosystème Hadoop & NOSQL et ses ramifications, avec la mutation vers l'analytique temps réel qui apparaît actuellement (Apache Drill, Google Dremel, etc.).

Des retours d'expérience, des cas issus de « la vraie vie ». On commence à voir des choses intéressantes depuis quelques mois, et cela dans tous les secteurs d'activités...



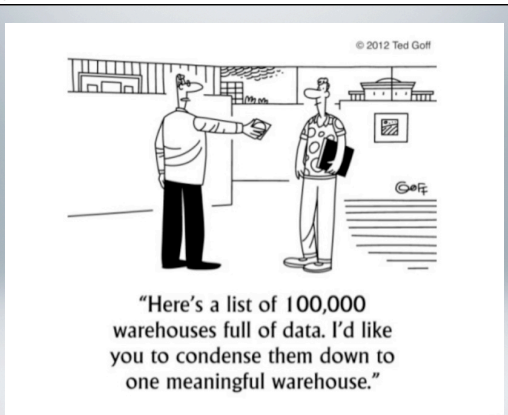


Affini-TECH

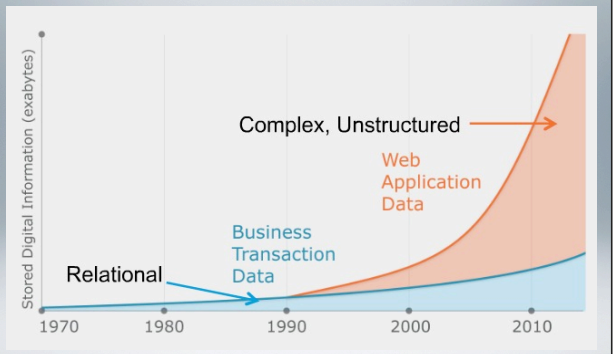
AGENDA

- BigData
- Hadoop & Datawarehouses
- Evolutions
- Performances
- Cas d'utilisation

Affini-TECH



Affini-TECH



Affini-TECH

Affini-TECH

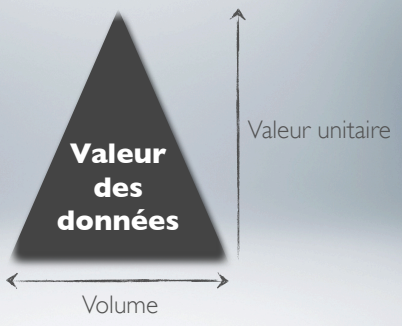
LES 4 V DU BIGDATA

Affini-TECH

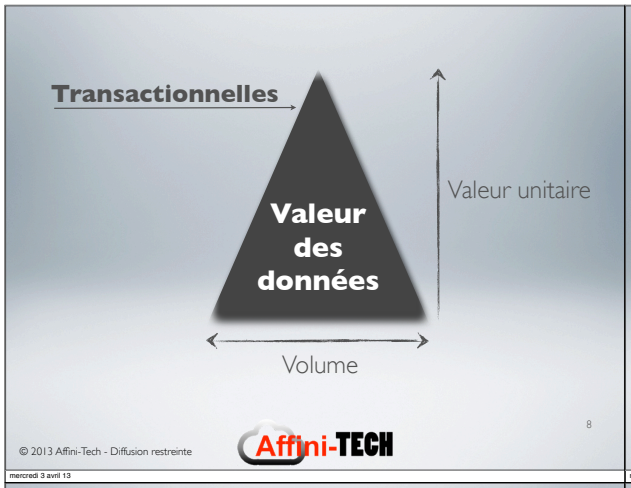
LES 4 V DU BIGDATA

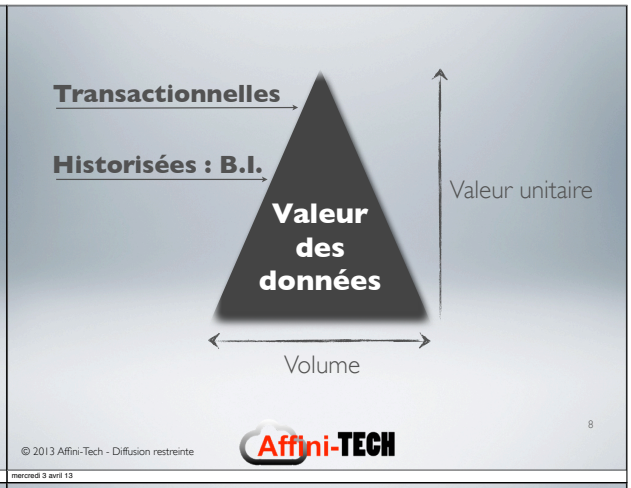
- **Volume** : les technologies actuelles sont inadaptées à cette croissance effrénée.
- **Variété** : l'entreprise est confrontée à des données non structurées : emails, web, réseau sociaux, son, image, video...
- **Vélocité** : L'accès et le partage des données doit se faire en temps réel.
- **Variabilité** : On ne sait pas prévoir l'évolution des types de données

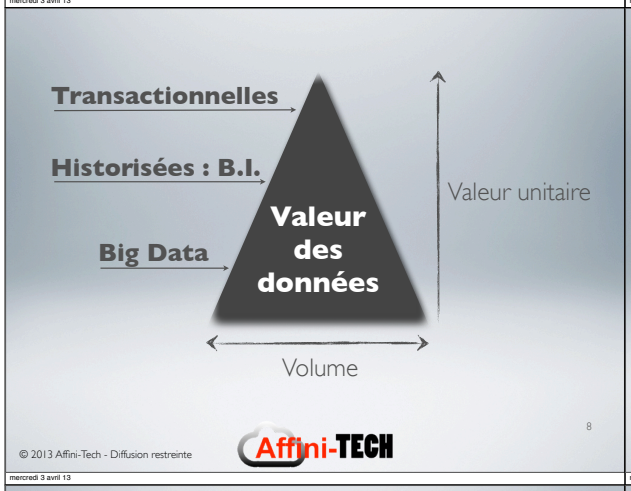
Affini-TECH

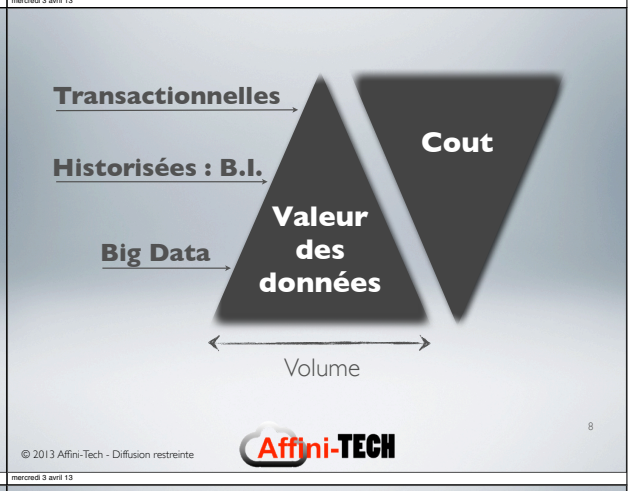


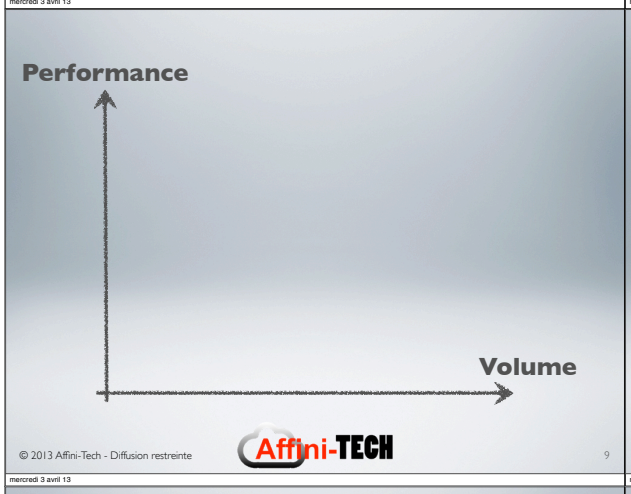
Affini-TECH

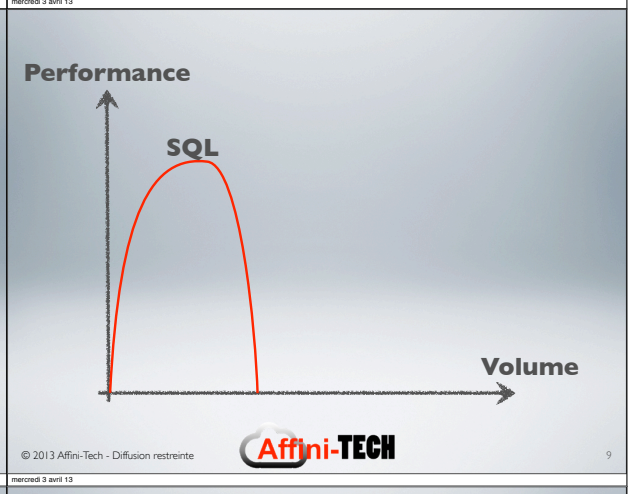

Transactionnelles
Valeur des données
Valeur unitaire
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
8
mercredi 3 avril 13

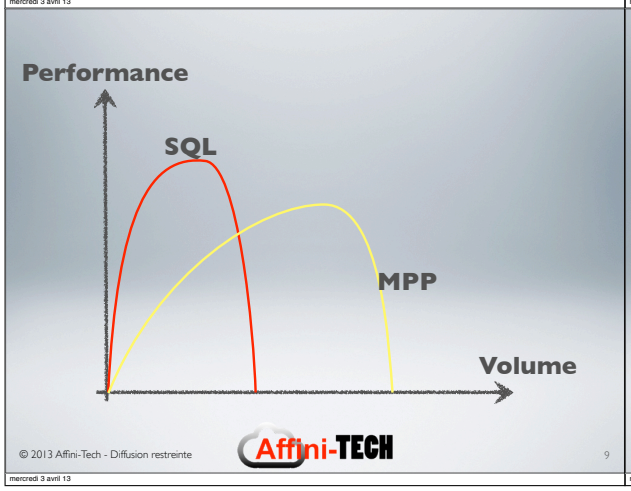

Transactionnelles
Historisées : B.I.
Valeur des données
Valeur unitaire
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
8
mercredi 3 avril 13

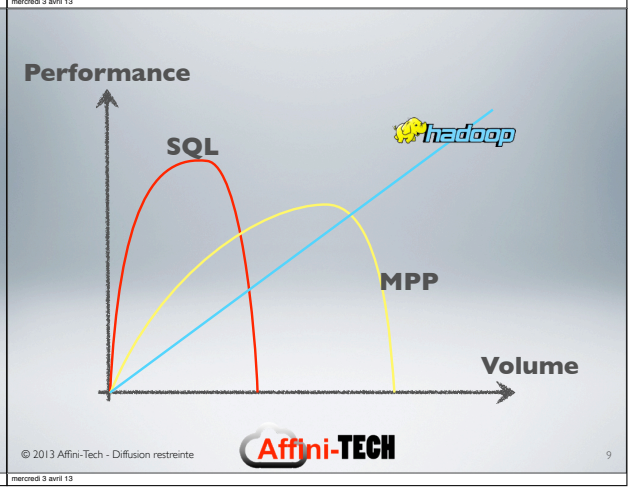

Transactionnelles
Historisées : B.I.
Big Data
Valeur des données
Valeur unitaire
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
8
mercredi 3 avril 13

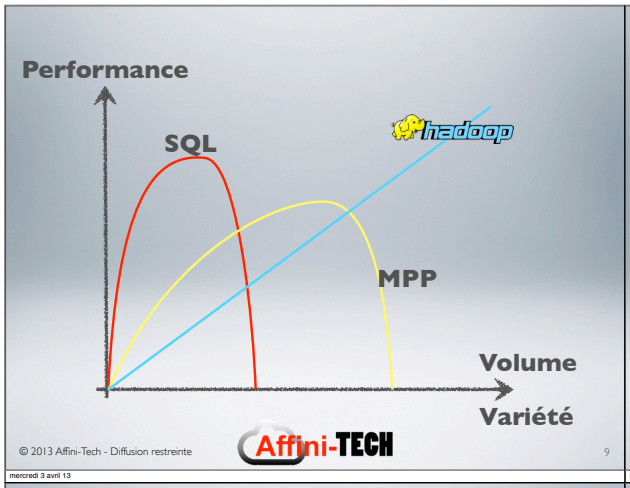

Transactionnelles
Historisées : B.I.
Big Data
Cout
Valeur des données
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
8
mercredi 3 avril 13

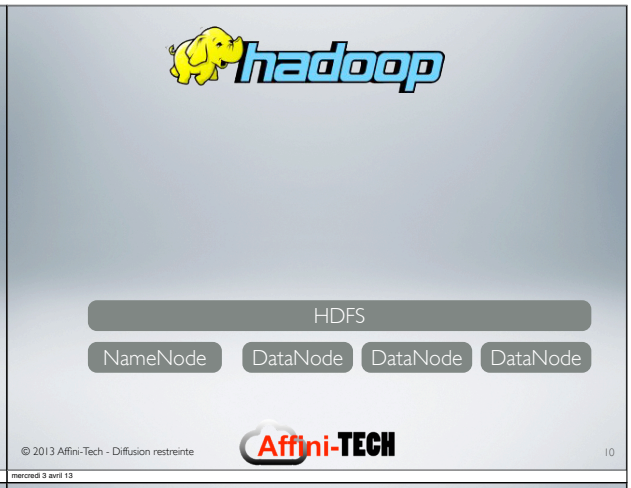

Performance
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
9
mercredi 3 avril 13

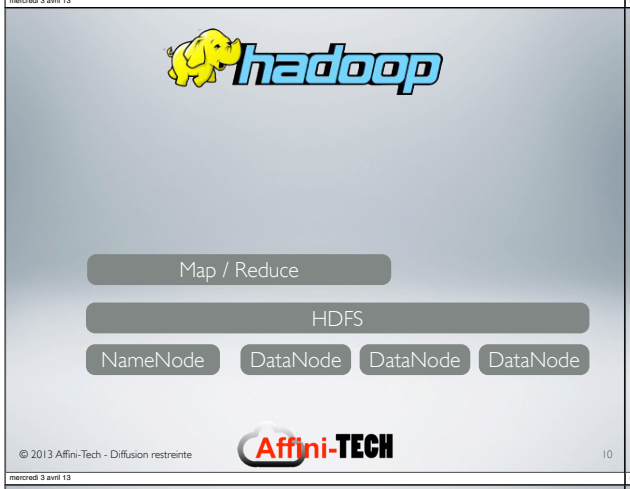

Performance
SQL
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
9
mercredi 3 avril 13

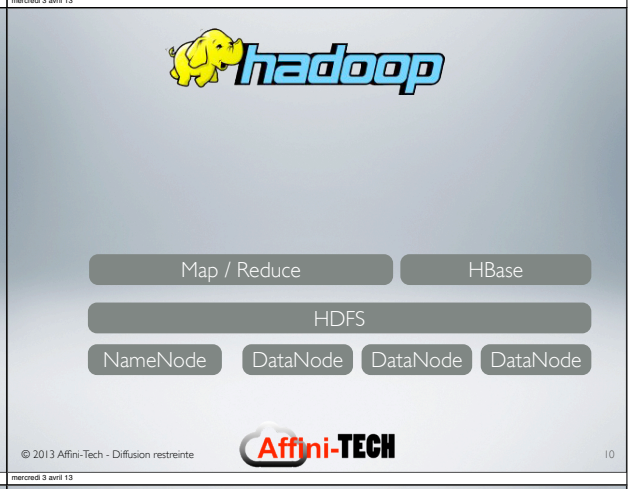

Performance
SQL
MPP
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
9
mercredi 3 avril 13

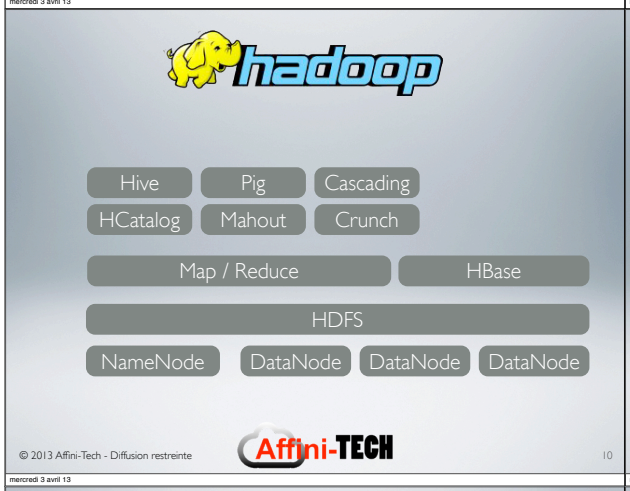

Performance
SQL
hadoop
MPP
Volume
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
9
mercredi 3 avril 13

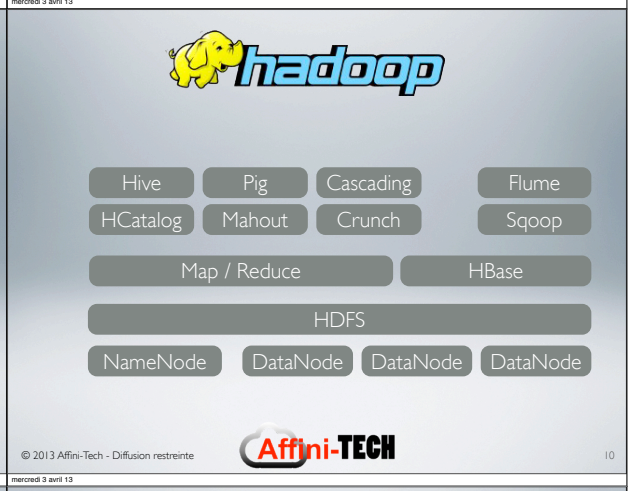

Performance
SQL
hadoop
MPP
Volume
Variété
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
9
mercredi 3 avril 13

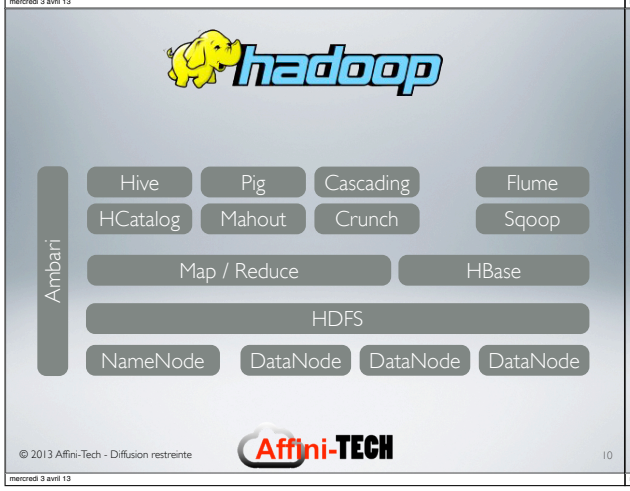

hadoop
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13

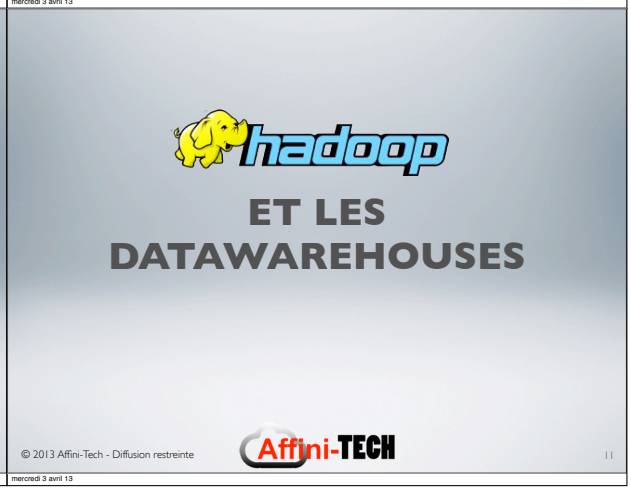

hadoop
Map / Reduce
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13


hadoop
Map / Reduce
HBase
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13


hadoop
Hive
Pig
Cascading
HCatalog
Mahout
Crunch
Map / Reduce
HBase
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13


hadoop
Hive
Pig
Cascading
Flume
HCatalog
Mahout
Crunch
Sqoop
Map / Reduce
HBase
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13


hadoop
Ambari
Hive
Pig
Cascading
Flume
HCatalog
Mahout
Crunch
Sqoop
Map / Reduce
HBase
HDFS
NameNode
DataNode
DataNode
DataNode
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
10
mercredi 3 avril 13


hadoop
ET LES
DATAWAREHOUSES
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
11
mercredi 3 avril 13

B.I. TRADITIONNELLE

Transactionnel

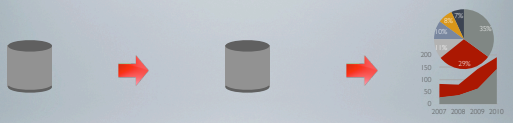
Affini-TECH

B.I. TRADITIONNELLE

Transactionnel DataWarehouse

Affini-TECH

B.I. TRADITIONNELLE



Transactionnel DataWarehouse BI Applications

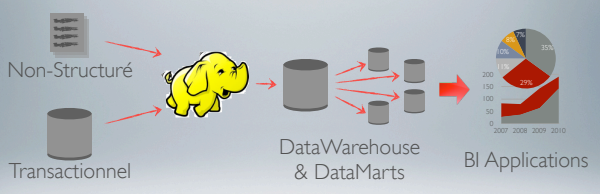
Affini-TECH

B.I. TRADITIONNELLE

Transactionnel DataWarehouse & DataMarts BI Applications

Affini-TECH

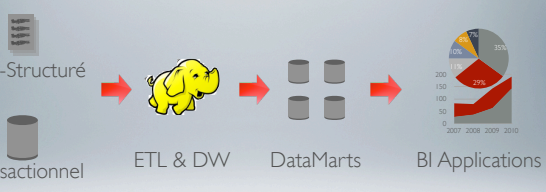
hadoop : ETL++



Non-Structuré Transactionnel DataWarehouse & DataMarts BI Applications

Affini-TECH

hadoop : ETL & DW



Non-Structuré Transactionnel ETL & DW DataMarts BI Applications

Affini-TECH

hadoop : EDW

Non-Structuré Transactionnel ETL & DW & DataMarts BI Applications

Affini-TECH

EVOLUTIONS

Différentes Workloads

Map / Reduce ne suffit plus

Productivité du développeur

Ouverture de l'écosystème

Performances

Affini-TECH

TYPES DE WORKLOADS

	Batch
Latence	Minutes à Heures
Volume	To à Po
Modèle	Map / Reduce
Utilisateurs	Développeurs

Affini-TECH

TYPES DE WORKLOADS

	Batch	Stream
Latence	Minutes à Heures	Continu
Volume	To à Po	Flux continu
Modèle	Map / Reduce	DAG
Utilisateurs	Développeurs	Développeurs

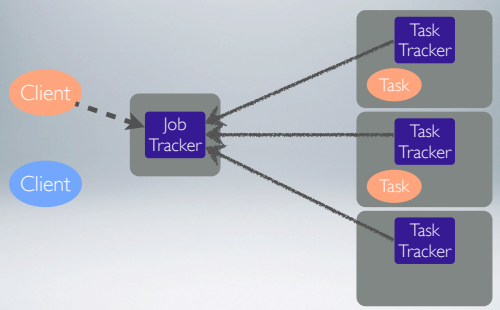
Affini-TECH

TYPES DE WORKLOADS

	Batch	Stream	Interactif
Latence	Minutes à Heures	Continu	Millisecondes à Minutes
Volume	To à Po	Flux continu	Go à Po
Modèle	Map / Reduce	DAG	Requêtes SQL
Utilisateurs	Développeurs	Développeurs	Analystes

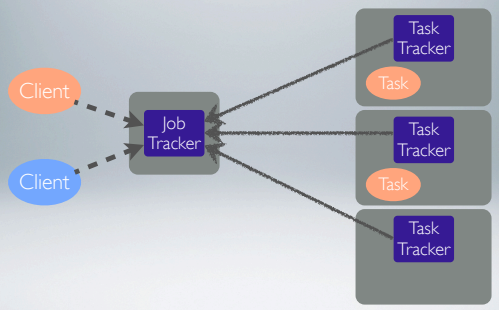
Affini-TECH

HADOOP 1 : MAP / REDUCE



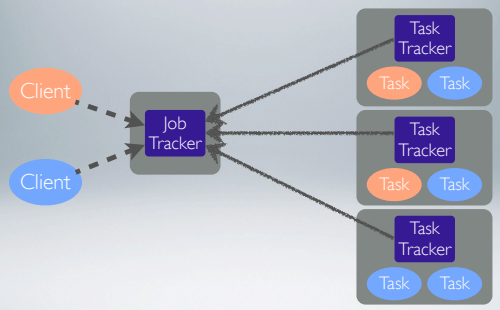
Affini-TECH

HADOOP 1 : MAP / REDUCE



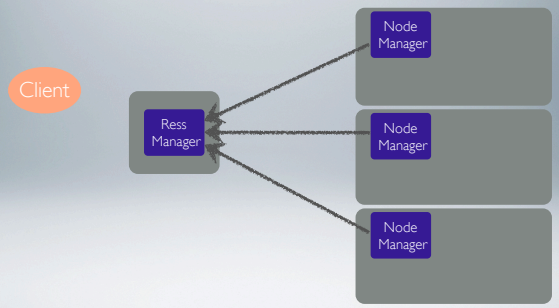
Affini-TECH

HADOOP 1 : MAP / REDUCE



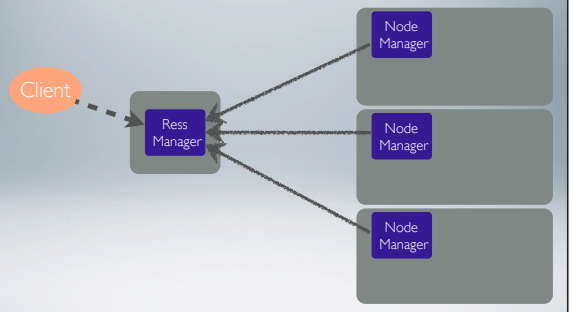
Affini-TECH

HADOOP 2 : YARN

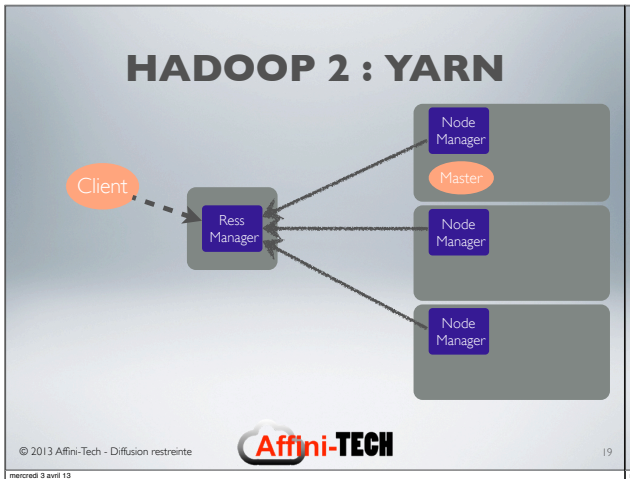


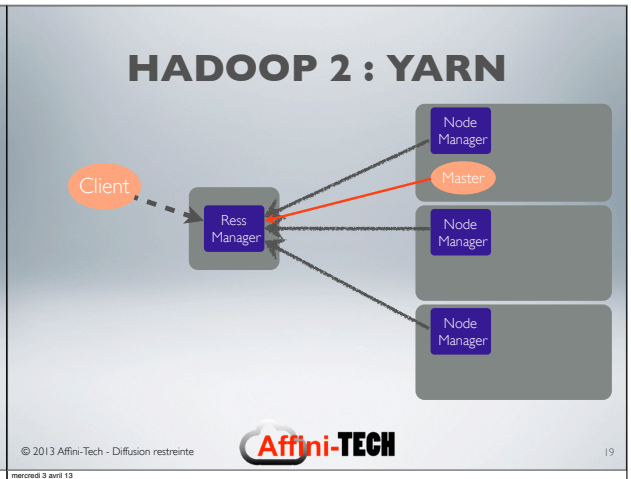
Affini-TECH

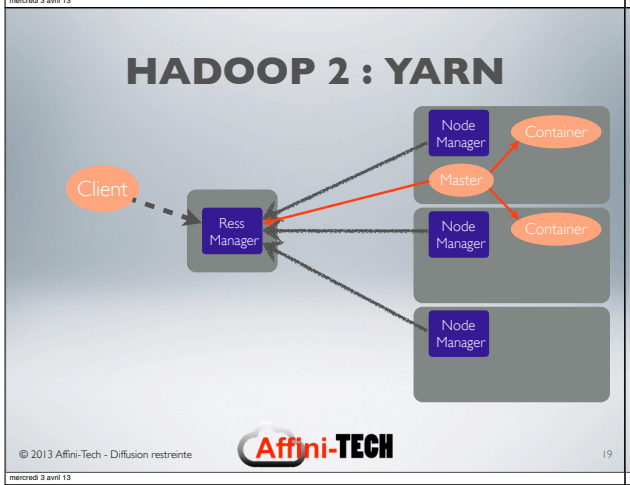
HADOOP 2 : YARN

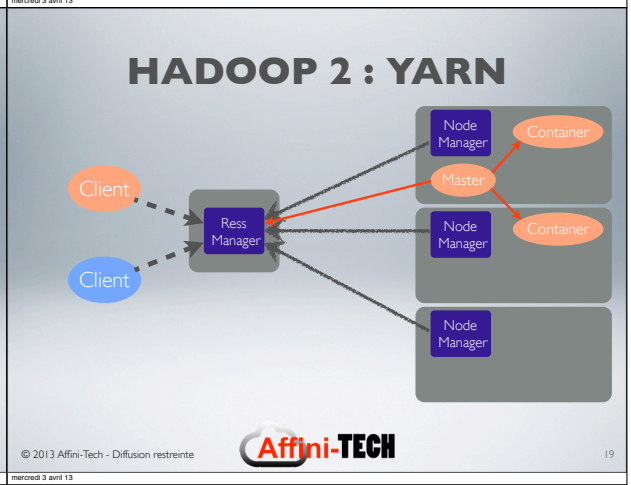


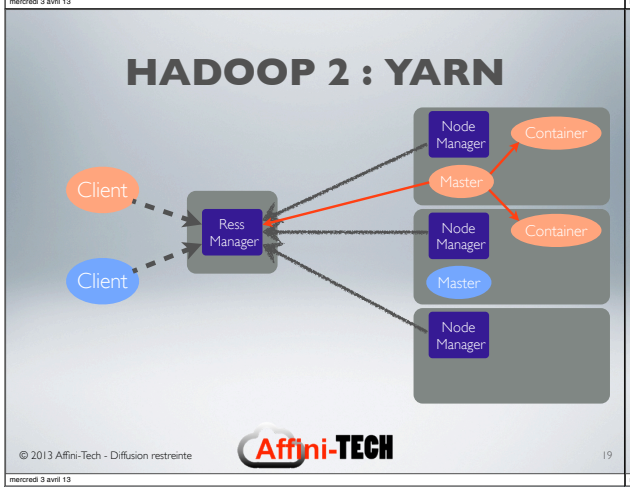
Affini-TECH

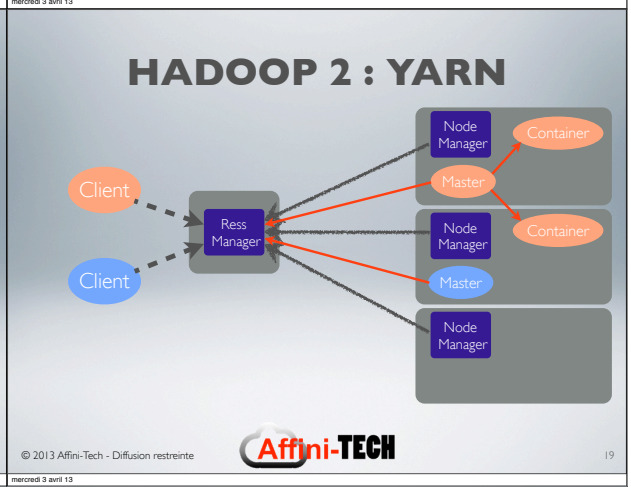

HADOOP 2 : YARN
Client
Ress Manager
Node Manager
Master
Node Manager
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13

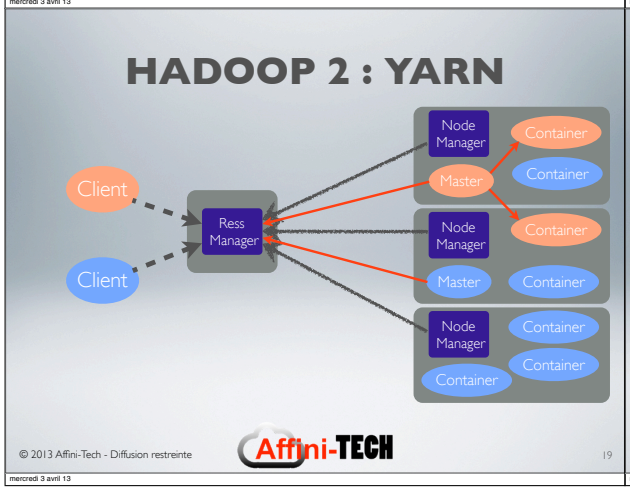

HADOOP 2 : YARN
Client
Ress Manager
Node Manager
Master
Node Manager
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13

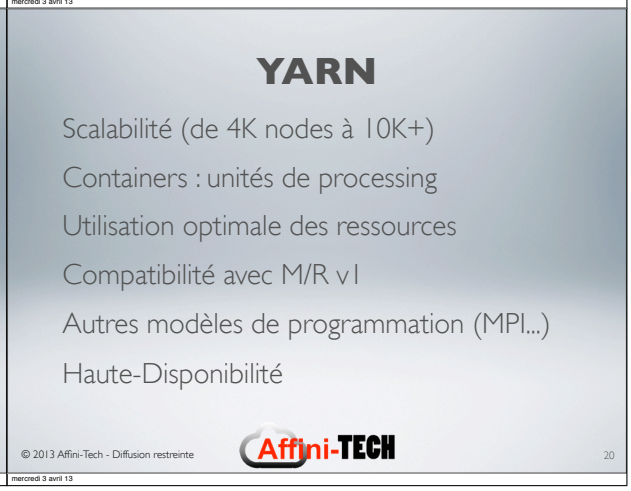

HADOOP 2 : YARN
Client
Ress Manager
Node Manager
Container
Master
Node Manager
Container
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13


HADOOP 2 : YARN
Client
Client
Ress Manager
Node Manager
Container
Master
Node Manager
Container
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13


HADOOP 2 : YARN
Client
Client
Ress Manager
Node Manager
Container
Master
Node Manager
Container
Master
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13


HADOOP 2 : YARN
Client
Client
Ress Manager
Node Manager
Container
Master
Node Manager
Container
Master
Node Manager
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13


HADOOP 2 : YARN
Client
Client
Ress Manager
Node Manager
Container
Master
Container
Node Manager
Container
Master
Container
Node Manager
Container
Container
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
19
mercredi 3 avril 13


YARN
Scalabilité (de 4K nodes à 10K+)
Containers : unités de processing
Utilisation optimale des ressources
Compatibilité avec M/R v1
Autres modèles de programmation (MPI...)
Haute-Disponibilité
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
20
mercredi 3 avril 13

PRODUCTIVITÉ DU DEVELOPPEUR

Map/Reduce est contraignant !

Alternatives masquant Map/Reduce :

- HIVE : SQL (+ interfaces JDBC)
- PIG : Séquences simples de transformation
- CASCADING : modèle de programmation simplifié pour tous les langages de la JVM

 Affini-TECH

OUVERTURE DE L'ÉCOSYSTEME

Possibilité de substituer des parties d'Hadoop par des codes extérieurs.

syncsort remplace le tri natif de Hadoop pour améliorer les performances.

Remplacement des connecteurs Hadoop par ceux d'ETL classiques du marché

 Affini-TECH

PERFORMANCES

Hybridation Hadoop/RDBMS

Impala : I/O directes & Bypass HDFS

Tez : Réduction de la latence

Spark : Map/Reduce in-memory

 Affini-TECH

HADOOP + RDBMS

Exporter les résultats de requêtes Hadoop vers un SGBD ou un appliance MPP

Mixer un SGBD classique et un stockage Hadoop
Le SGBD cache les données...
Hadapt, CitusDB, PivotalHD, Microsoft Polybase

 Affini-TECH

CLOUDERA IMPALA

Projet propriétaire de Cloudera

Fonctionnement proche des moteurs MPP & conserve un socle Hadoop

Lecture directe des blocs sur disques

Format colonne

Etend les interfaces de Hive/SQL

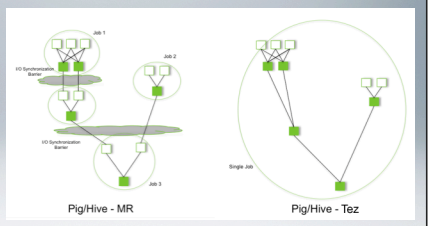
 Affini-TECH

APACHE TEZ & STINGER

Supprimer les I/O intermédiaires

Performances x45

Générique M/R



 Affini-TECH

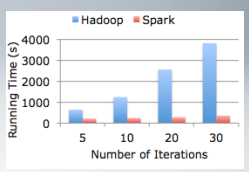
SPARK & SHARK

Spark : Implémentation de M/R en mémoire.
Structures de données distribuées.

Performances sur les iterations : Machine-Learning

Shark offre une compatibilité Hive/SQL

Un projet de amplab



 Affini-TECH

CAS D'UTILISATION

Facebook

Linkedin

Comscore

Voyages SNCF

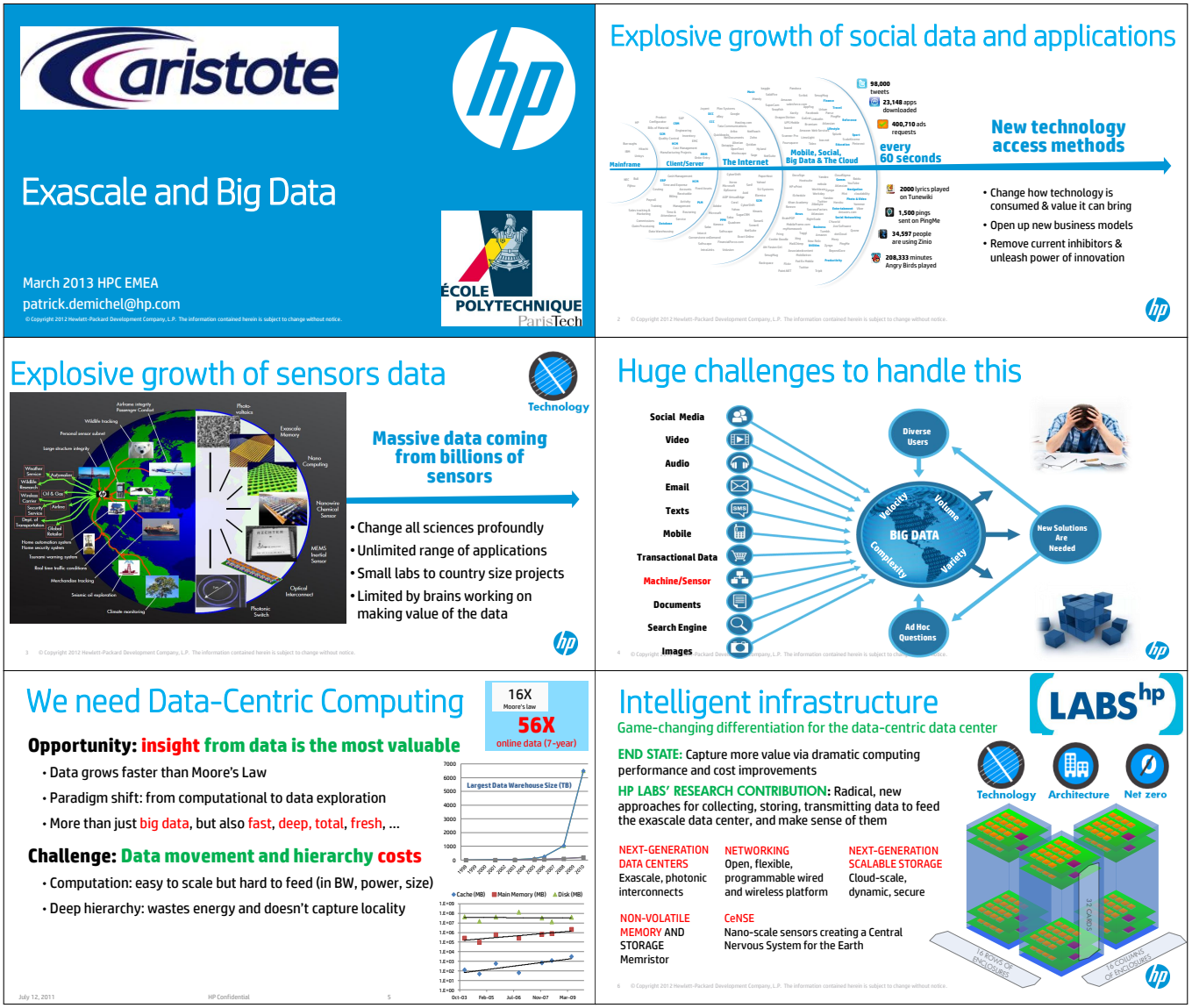
 Affini-TECH


MERCI !
Vincent Heuschling
Gsm : 06 61 88 76 71
Email : vhe@affini-tech.com
Web : http://www.affini-tech.com
Twitter : @affinitech & @vhe74
© 2013 Affini-Tech - Diffusion restreinte
Affini-TECH
29
mercredi 3 avril 13

3.3 Patrick Demichel (HP)

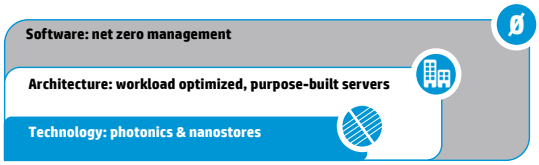
Évolutions technologiques pour le `Big Data`

- pourquoi un grand nombre de data, impact sur les architectures ;
- tour de piste des technologies *disruptives* ;
- focus sur les mémoires non volatiles de grosse capacité ;
- focus sur les communications photoniques.



The future Data-Centric Data Center

Electrons compute, photons communicate, ions store … at net-zero energy

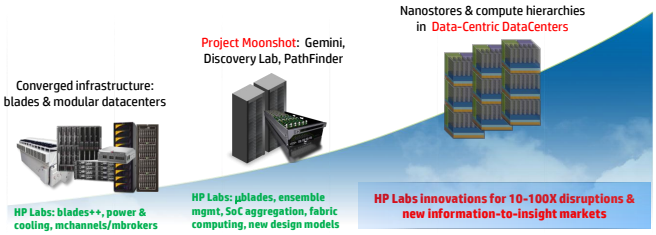


Holistic redesign for big impact – a "perfect storm" confluence of technology and application trends

Project and roadmap

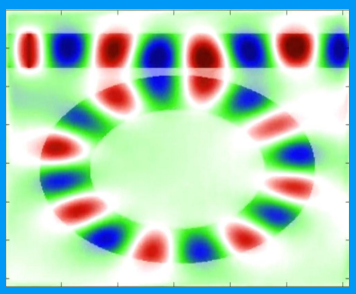
Holistic, systematic & step-wise roadmap to revolutionary impact

Architecture



Photonic

Technology



HP Photonics

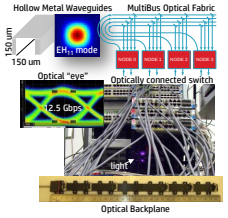
for faster data transmission and lower energy

Technology

Transmit data using light for 30-fold more bandwidth at one-tenth the energy

- **Bandwidth scaling**
  - 30x improvement over copper
- **Lower Power**
  - 10x improvement over copper links
  - Improved airflow & cooling
- **Equivalent cost**

All communications will be optical



Microrings

Full link configuration

Advantages

Modulators wavelength specific, no additional mux

Same ring structure used for drop filters

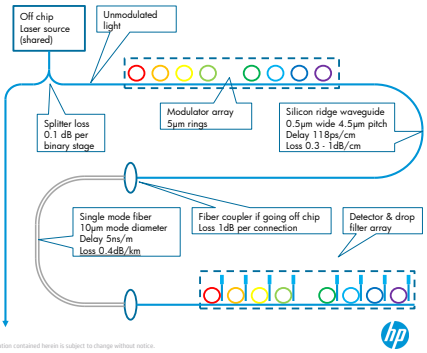
Loss budget dominated by cost

Up to 64 wavelengths

Outstanding issues

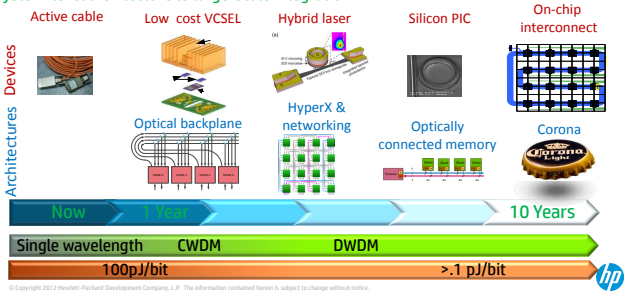
Ring tuning

Thermal stability

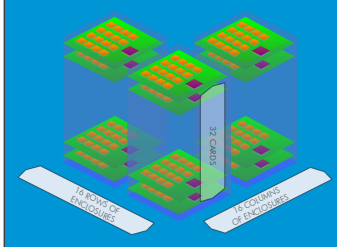


HP photonics technologies

System-level architecture to large-scale integration

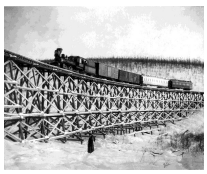


Photonic fabric



Topologies

There is not and never will be one "Best Topology"



PHOTONICS

High cost of end points

Long spans

Incredibly elegant

ELECTRONICS

Short "hops"

Cheap basic materials

Gets the job done

Future Router Investigation[1]

Rational – ideal pull through for integrated photonics

Exploits high bandwidth density and distance independence

Three designs options:

All electronic

Electronic with integrated photonic I/O

Photonic crossbar with photonic I/O

Validate for:

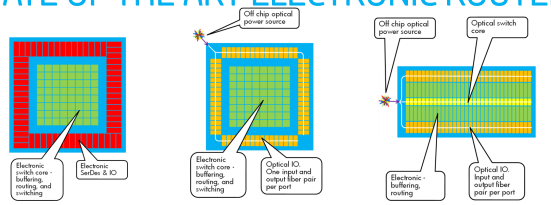
Performance

Power

Packaging

1. Binkert, N, et al."The role of optics in future high radix switch design." ISCA 2011

STATE OF THE ART ELECTRONIC ROUTER



Basic electronic router design uses hierarchical structure

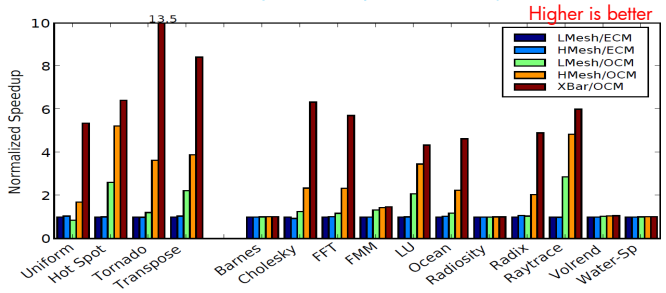
Similar to YARC but with support for variable length packets (64bytes – 8KBytes)

64, 100 & 144 port variants studied

Modelled using CACTI for power and M5 for simulation

Performance (LMesh/ECM = 1)

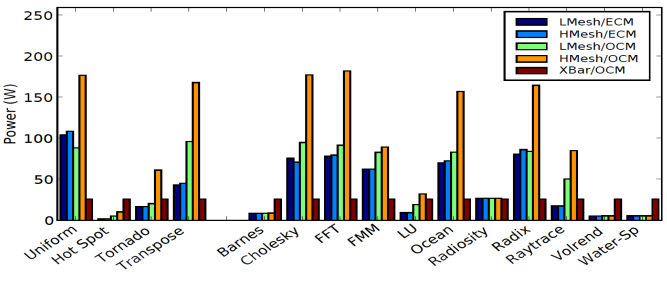
Higher is better



Applications that don't fit in cache show 4-6X improvements with Xbar

On-chip Network Power

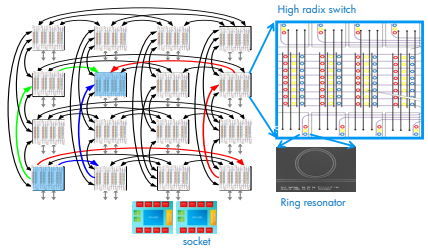
Lower is better



Optics can reduce network power of applications that don't fit in cache by 6X

HyperX[1] networks

Fully connected sub-networks in multiple dimensions

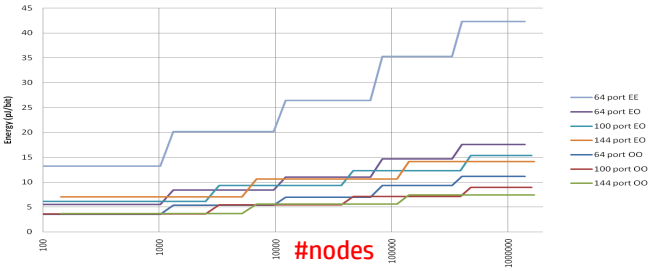


- Superset of "flattened butterfly" networks and hypercubes
- Fully connected networks offer lowest hop count but limited scalability
- Multiple dimensions increase scalability at the expense of hop count
- Many alternate paths with one or more additional hop
- Non-minimal routes required for full bisection bandwidth

1: "HyperX: Topology, Routing, and Packaging of Efficient Large-Scale Networks" Ahn et al., Supercomputing 2009

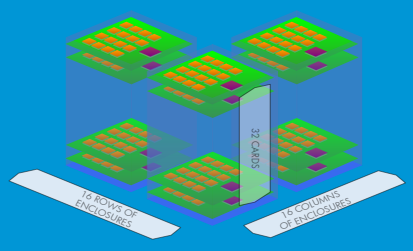
NETWORK POWER – LARGE HYPERX

Highest radix OO are the best options for large scale systems



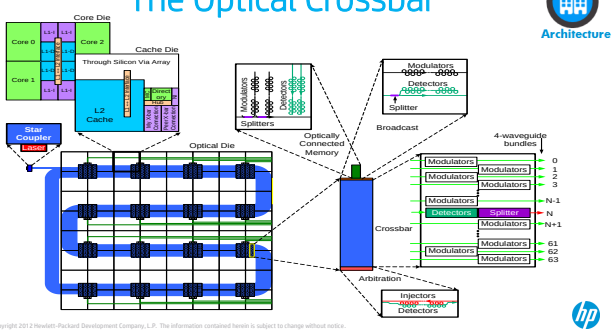
22nm process node, 320Gbps ports

PUTTING IT ALL TOGETHER



The Optical Crossbar

Architecture



AN EXASCALE COMPUTE NODE

Architecture

Single socket highly parallel CPU/GPU

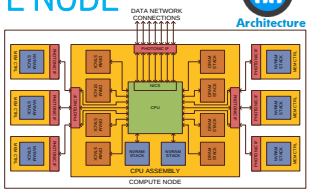
- Photonics interconnect for all connections off compute complex
- Multiple photonic links
- 8 x 0.4Tbps links
- Scalable compute communication ratio

Tightly coupled DRAM

- Direct stacking or high performance substrate
- Pull through for 3D stacking technologies

Use common photonic physical layer for "far" memory

Broadcast for memory scaling?



Node Performance Targets

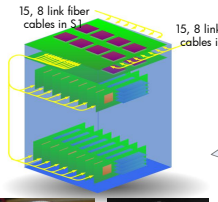
Node Peak Performance12-14Tflops

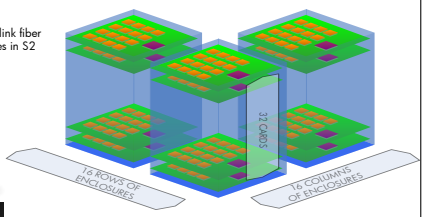
Memory BW >1Tbyte/s

Node Network BW 400Gbyte/s

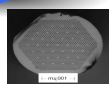
Power <200W

128K nodes system = 16*16*5PF = 1.3EF





72 way parallel fiber ribbon



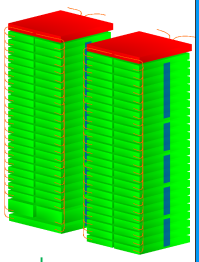
37 core photonic crystal fiber (PCF)

16 x 16 array of enclosures

Total fiber ribbons:

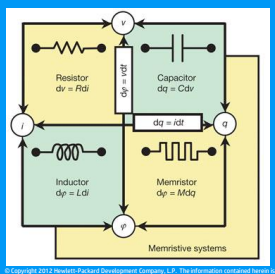
$$\frac{16^2 \times (15+15) \times 2}{2} = 7680$$

Conclusions



- Integrated photonics has the potential to:
  - Dramatically improve memory bandwidth
  - Significantly improve many-core performance
  - Reduce power
  - Simplify programming
  - All at the same time!
- Near term applications such as optical buses
  - Add significant system flexibility
  - Save latency and power
- Longer term give opportunity to rethink system arch
  - New architectures & flexibility (e.g., optical buses)
  - Disaggregation and dematerialization enablement

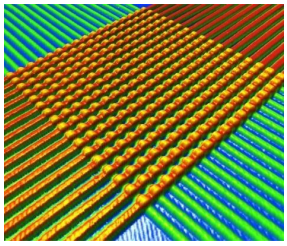
Non Volatile Memory : Memristor



Non Volatile Memory : Memristor

Non Volatile Memories are fundamental for Exascale success

- Power ; performance ; architecture flexibility ; fault tolerance
- Programmability ; capacity and cost
- Interesting properties:
  - High density $4F^2$, stackable
  - Low cost
  - Low power : pJ/bit
  - High speed : ns
  - High endurance : >> 10^10
  - High retention : 10+ years
  - Reconfigurable architectures
  - Multiple optimized variants
  - Long term roadmap : post Moore
  - Can do logic « better transistors »
  - Can do neuristor



Technology Attributes

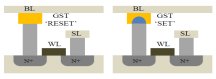
- Scaling down to less than 10 nm width per cell
  - ~ 32 Gbyte/cm^2/layer by 2018
- Scaling up to multiple (≥ 8) layers on chip
  - ~ 0.25 Tbyte/cm^2/chip by 2018
- Truly nonvolatile – many, many years
- Random Access
- Fast cell write and erase (~ nanosec)
- Low energy cell write and erase (~ picoJ)
- Good to excellent endurance (> 10^{10} cycles)
  - Still counting – goal is to exceed 10^{18} cycles

Technologies for Check-point Restart

www.nd.edu/~rich/SC09/tut157/SC2009_Jouppi_Xie_Tutorial_Final.pdf

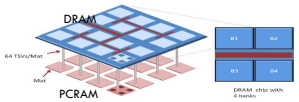
Architecture

PCRAM



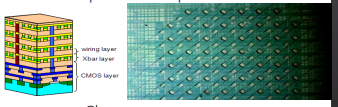
The schematic view of a PCRAM cell with NMOS access transistor (BL=Bitline, WL=Wordline, SL=Sourceline)

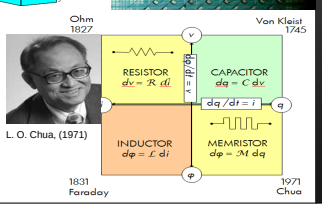
	HDD	NAND Flash	PCRAM
Taille cellule	-	4-6F^2	4-6F^2
Cycle lecture	~4ms	5us-50us	10ns-100ns
Cycle ecriture	~4ms	2ms-3ms	100-1000ns
Watt à arrêt	~1W	~0W	~0W
Endurance cycles	10^15	10^5	10^8

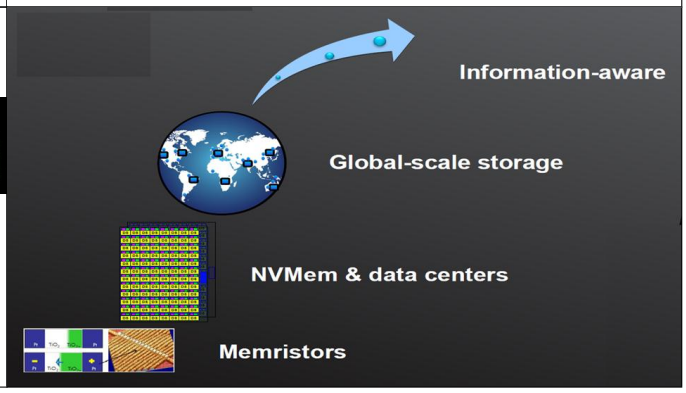


Memristor

CMOS chip avec des composants memrésistifs



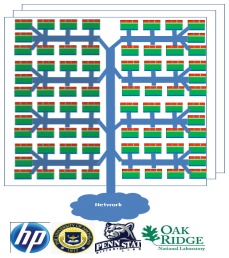




Blackcomb

Hardware-Software Co-design for Non-Volatile Memory in Exascale Systems
Opportunities go far beyond a plugin replacement for disk drives…

- New distributed computer architectures that address exascale resilience, energy, and performance requirements
  - replace mechanical-disk-based data-stores with energy-efficient non-volatile memories
  - explore opportunities for NVM memory, from plug-compatible replacement (like the NV DIMM, below) to radical, new data-centric compute hierarchy (nanostores)
  - place low power compute cores close to the data store
  - reduce number of levels in the memory hierarchy
- Adapt existing software systems to exploit this new capabilities



From microprocessors to nanostores for maximum efficiency

Game-changing differentiation for the data-centric data center

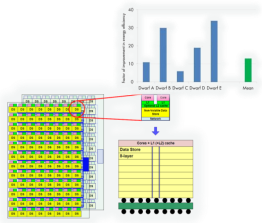
Technology

Enabled by HP Memristors technology,

HP Nanostores provide flat converged storage hierarchy with compute colocation for

10-100X better performance/watt

- **More efficient insight extraction from cold data**
- **Fast insights on hot data**



Global Scale Storage

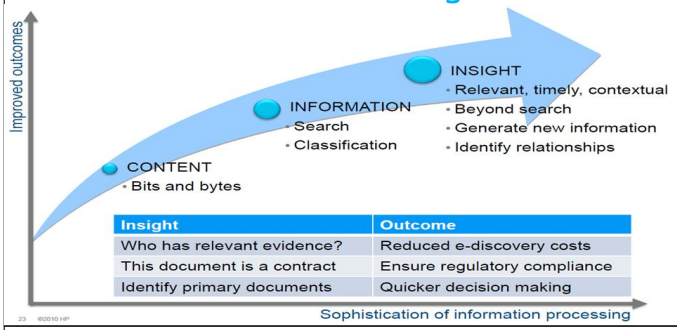
- Global infrastructure
- Global clients/applications
- Research challenges
  - Scalability
  - Availability
  - Low cost
  - Flexibility

erasure codes: Low cost availability

- Example: 4 data + 8 "parity" fragments, any 4 can recover

- Fault tolerance
  - Tolerates loss of one **entire** data center
  - Each data center independently tolerates any two disk failures
  - Eight disk failures tolerated across data centers
- Space efficient
  - Overhead of 3x replication with fault tolerance of 9x replication
  - Can tune the space efficiency-reliability trade off
- Costs
  - Computation for encode and decode
  - To recover on failed disk, 4 disks' worth of data must be read
- Tunable tradeoff between storage efficiency and fault tolerance

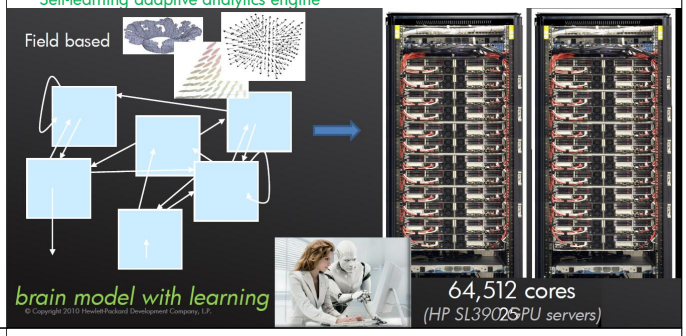
Vision : from content to insight



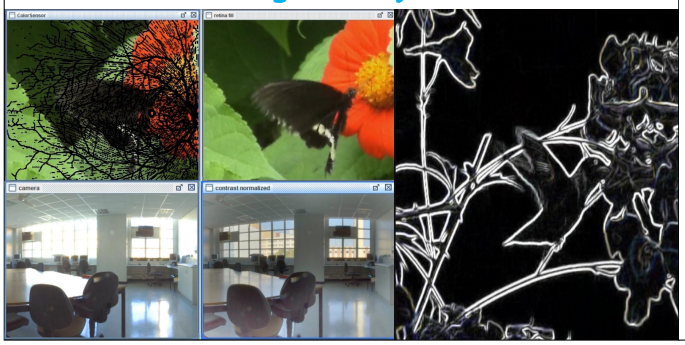
Insight	Outcome
Who has relevant evidence?	Reduced e-discovery costs
This document is a contract	Ensure regulatory compliance
Identify primary documents	Quicker decision making

Neuristors and cognitive systems

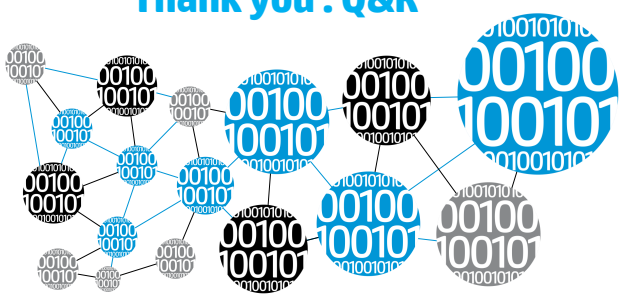
Self-learning adaptive analytics engine



Neuristors for cognitive systems



Thank you : Q&R

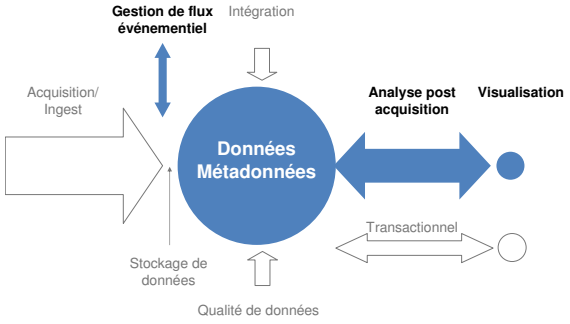


3.4 Guy Chesnot (SGI France)

In memory

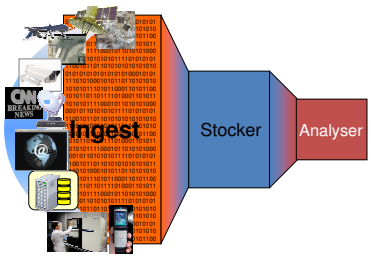
Le terme Big Data couvre à la fois un aspect quantité – de volumétrie, de vitesse d'arrivée des données et de multiplicité des types de données numériques – et un aspect algorithmique : comment retirer des informations de valeur de cette masse de données ? Cette présentation se concentre sur l'aspect analyse en différenciant deux classes de méthodes selon les questions posées et les sources de données. On aboutit ainsi à une dichotomie entre traitements distribués et traitements *In-memory*, distinction illustrée par des exemples industriels ou issus de la recherche.

Actions sur les Big Data (ce jour)

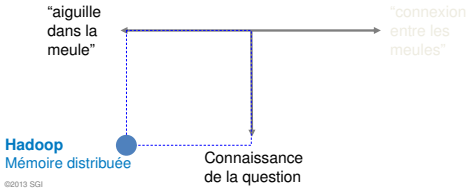


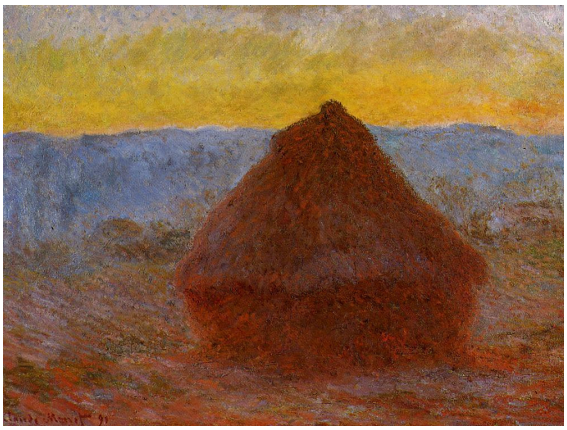
Big Data

Par jour : **Go → To → Po → Eo**

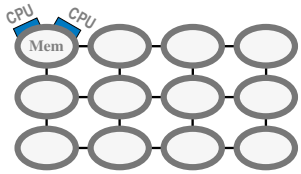


Analyse : architectures





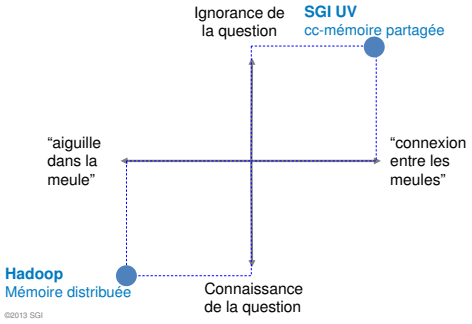
Architecture à mémoire distribuée: cluster Hadoop



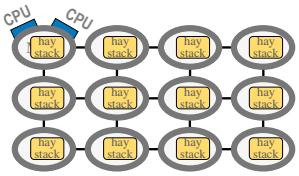
Gestion d'événements: analyse Big Data sans stockage de données!

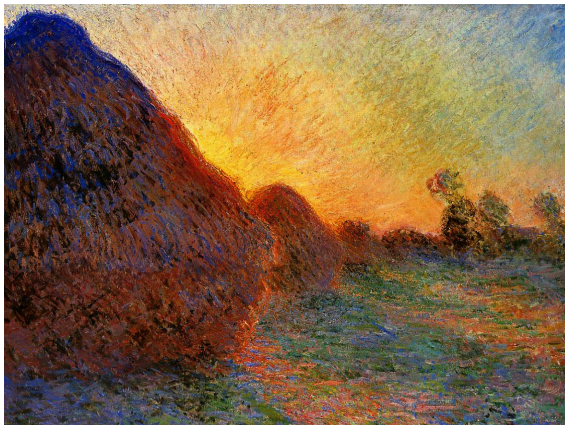
- Analyse d'événements de Set top box
- Latence temps réel
  - Analyse
  - Performance
  - Réponse aux incidents

Analyse : architectures

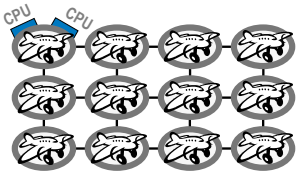


Architecture à mémoire distribuée: cluster Hadoop





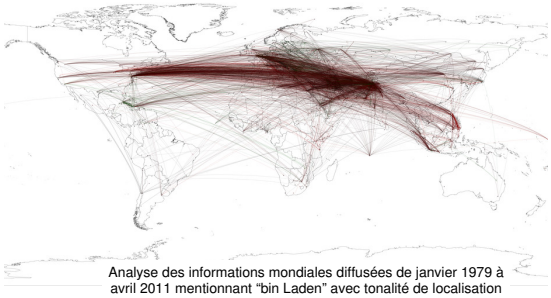
Nombreuses meules de foin (suite)



Analyse Big Data sans Hadoop connexion entre les meules

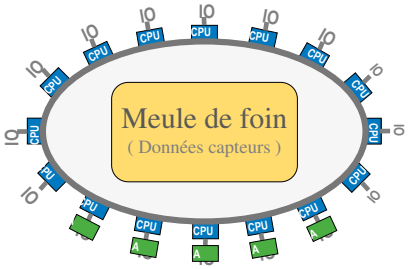
Connexion de tous les endroits cités la même année. English Wikipedia 2012.

Connexion entre les meules (suite)



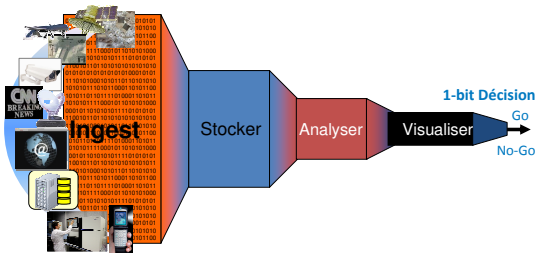
Analyse des informations mondiales diffusées de janvier 1979 à avril 2011 mentionnant “bin Laden” avec tonalité de localisation

Architecture à mémoire partagée: SGI UV



Big Data to Decision

Par jour : **Go → To → Po → Eo**



Décision oui ou non Gestion d’événements et analyse

- Détection de fraude

eBay PayPal™

- Serveur intégré

WHAT KIND OF VOLUME?

10 million+ PayPal logins / day.

13 million financial transactions / day.

300 variables calculated per event for some models.

~4 Billion inserts / day.

~8 Billion selects / day.

PayPal™

http://www.hpcuserforum.com/presentations/dearborn2012/Arno_Kolster_PayPal.pdf

Décision sans connaissance de la question

- Détection de fraude

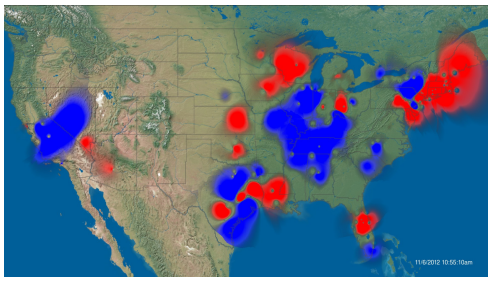
INPS
Istituto Nazionale della Previdenza Sociale

- Serveur à grande mémoire

Analyse Big Data sans Hadoop: Twitter heartbeat

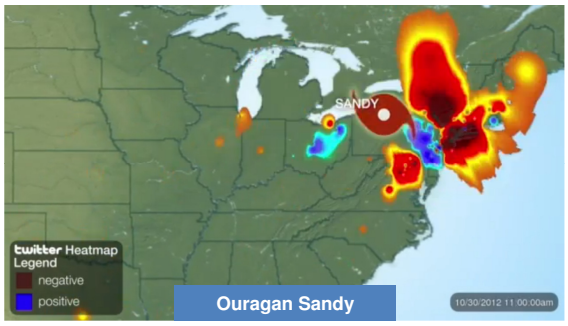
- Analyse géographique de Tweets sur UV 2000
- Analyse en temps réel de 10% des tweets quotidiens (soit 50 millions)
  - Assignation géographique selon GPS & le texte lui-même
  - Détection de la tonalité: données contextuelles (ton, géographie)
  - => visualisation cartographique de la densité et de la tonalité des conversations
- Sur SGI UV
  - Ingestion et traitement
  - Permet d'afficher une carte par seconde
  - Deux événements récents: élection Obama/Romney et ouragan Sandy

Analyse + visualisation + décision



Analyse Big Data sans Hadoop: Twitter heartbeat (suite 1)

Analyse Big Data sans Hadoop: Twitter heartbeat (suite 2)



3.5 Patrick Fuhrmann (Desy)

dCache un système de gestion de données réparties

Démarré il y a quelques années au centre du Desy à Hamburg, dCache s'est voulu être un système réparti de dépôts de données, de gestion de réplication et de gestion de hiérarchie. Dcache est le résultat d'une collaboration entre le Desy, FERMiLab et le NelC des pays nordiques, dans le cadre d'une initiative Européenne : EMI. Il a ensuite été choisi par le LHC du CERN, lui apportant ainsi la reconnaissance de sa maturité. Principalement utilisé en Europe du nord dCache étend sa zone d'influence marqué notamment par un *workshop* le 17 Mars à Tapei. Patrick Fuhrman présentera les spécifités techniques qui font le succès de dCache.

dCache.org

dCache, un système de gestion de données réparties

Mar 27, 2013 a la Séminaire Aristote
Patrick Fuhrmann

dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 1

dCache.org

Preview

- Some dCache project stuff
  - Funding, partners, deployments
- Software design and features
  - Modules and message passing
  - Namespace and physical location
  - Plug-in services
- Project objectives and consequences
  - Committed to standards
  - Benefits of collaborations
- The dCache labs

dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 2

dCache.org

The project … stuff

dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 3

dCache.org

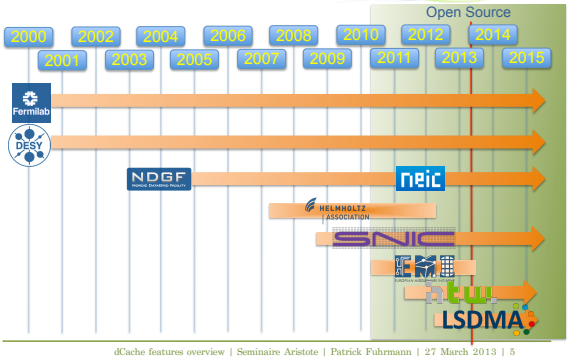
Projects and funding

dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 4

dCache.org

Partners and funding

dCache project timeline



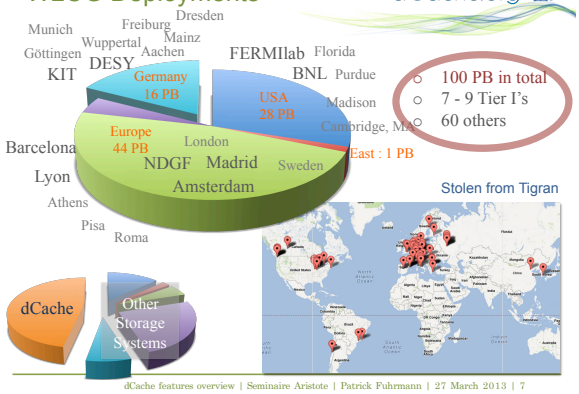
dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 5

dCache.org

Deployments

dCache features overview | Seminaire Aristote | Patrick Fuhrmann | 27 March 2013 | 6

WLCG Deployments



Other communities

Stolen from Paul

Most important for sustainability

For all major partners, dCache is a strategic system, running in production.

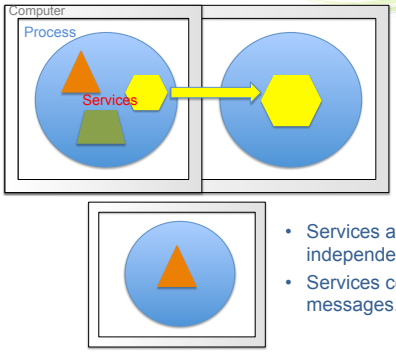
And now for something completely different

Software design and features

Design #1

Service Modules & Message Passing

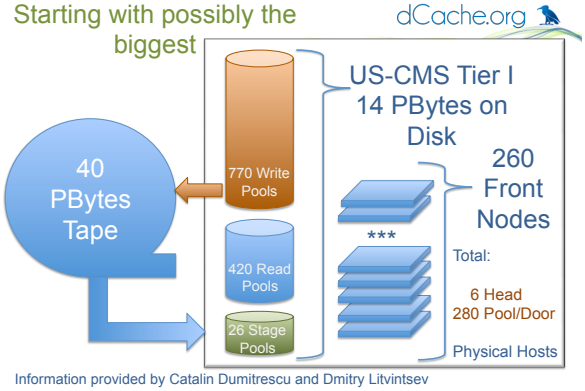
Scale-out Design



- Services are location independent.
- Services communicate via messages.

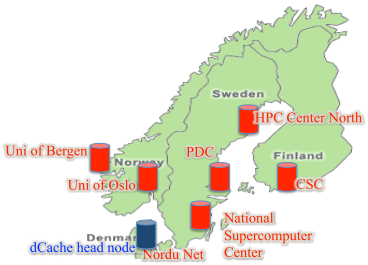
Resulting in …. Fits all sizes

Starting with possibly the biggest



Information provided by Catalin Dumitrescu and Dmitry Litvintsev

To certainly the most widespread



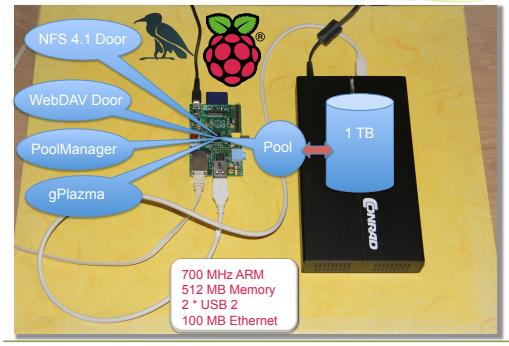
4 Countries

One dCache

Slide stolen from Mattias Wadenstein, NDGF

To very likely the smallest

One Machine – One Process

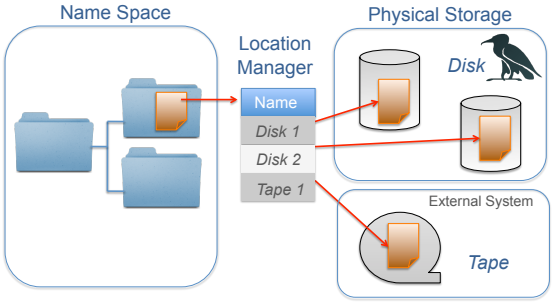


Design #2

Namespace – Physical Storage separation

Design

Namespace – Storage separation



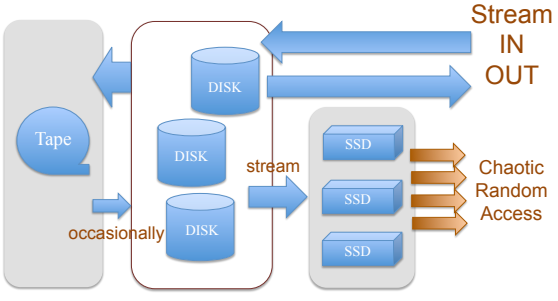
Resulting in …. Replica Management

Replica Management

- Hot Spot detection
  - Files are copied from 'hot' to 'cold' pools
- Multi Media Support
  - File location is based on access profile and storage media type/properties
    - Fast streaming from spinning disks
    - Fast random I/O from SSD's
- Migration Module(s)
  - Files can be manually/automatically moved or copied between pools.
  - Rebalancing of data after adding new (empty) pools.
  - Decommission pools.
- Resilient Manager
  - Keeps max 'n' min 'm' copies of a file on different machines.
  - System resilient against pool failures.
- Tertiary System connectivity (Tape systems)
  - Data is automatically migrating to tape.
  - Data is restored from tape if no longer on disk

e.g: File location management

Analysis



Design #3

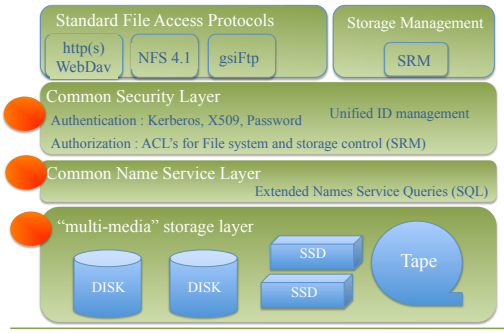
Services allow plug-ins

dCache.org

Resulting in … customizable behavior

Plug-in Facility

dCache.org



Plug-in Facility

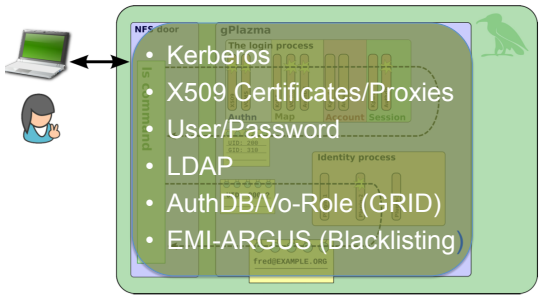
dCache.org

- gPlazma / Authentication system
  - Authentication
  - Mapping (user names and UID/GID)
  - Actually in the door:
    - LFN to PFN mapping for CMS and Atlas
- Name space provider (PNFS -> chimera)
- File System back end
- File distribution / reshuffling system

gPlazma plug-ins (e.g. NFS4.1)

dCache.org

Slide stolen from Paul Millar



dCache.org

Now … about some project objectives

dCache.org

Objective #1

Committed to standards

Resulting in …

dCache.org

- Support of
  - GLUE 2
  - SRM
  - WebDAV
  - NFS 4.1 / pNFS
  - The Storage Accounting Record (StAR)
- Working on Cloud protocols

Makes dCache an Open Source competitor to expensive industry solutions and attracts non WLCG communities.

dCache.org

Objective #2

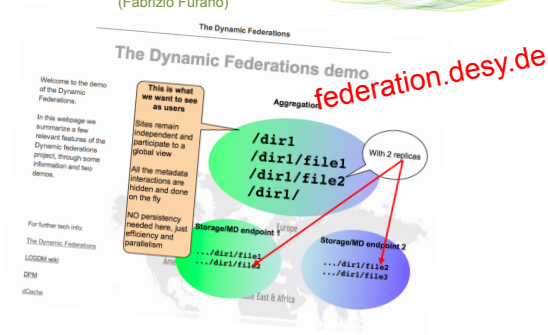
We believe in the power of collaborations

Resulting in

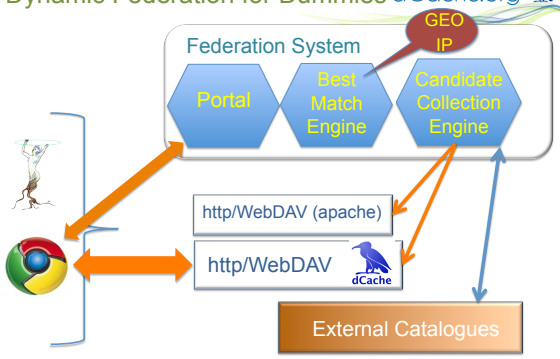
- European Middleware Initiative (EMI)
  - Funding for very interesting development
  - Learning about the storage needs of non HEP communIties
- CERN DM
  - HTTP Dynamic Federation
- Globus Online
  - gridFTP and staging
- **L**arge **S**cale **D**ata **M**anagement and **A**nalysis
  - about 'federated identity and storage access'

Dynamic Federation

(Fabrizio Furano)



Dynamic Federation for Dummies



The dCache labs

dCache labs

- Cloud storage protocols S3 and CDMI
  - HTW Berlin student working on those
- Enhanced 3 Tier storage
  - e.g. scheduling of data location changes
  - Migration of data based on access count
- Small Files and tape migration
- Adopting more standard identity mechanisms, IdP (e.g. Shibboleth, OpenID)

Where to learn more about dCache ?

- One workshop per year in Europe.
- One dCache day during the GridKA school.
- First Asian Pacific dCache in Taipei (last week).

With participants from

- Australia
- Taiwan
- Thailand
- Japan
- India
- Germany

Summary

- dCache is a professional Open Source project, with a large developers base and significant community support.
- Funding is provided by a variety of sources.
- dCache is committed to standards
  - To ease customer acceptance for storage
  - Simplifies system administrators life.
- The dCache system evolves, following
  - Community requirements (SRM, GLUE2, StaR …)
  - Technology changes (NFS 4.1, SSD, Hadoop FS, …)

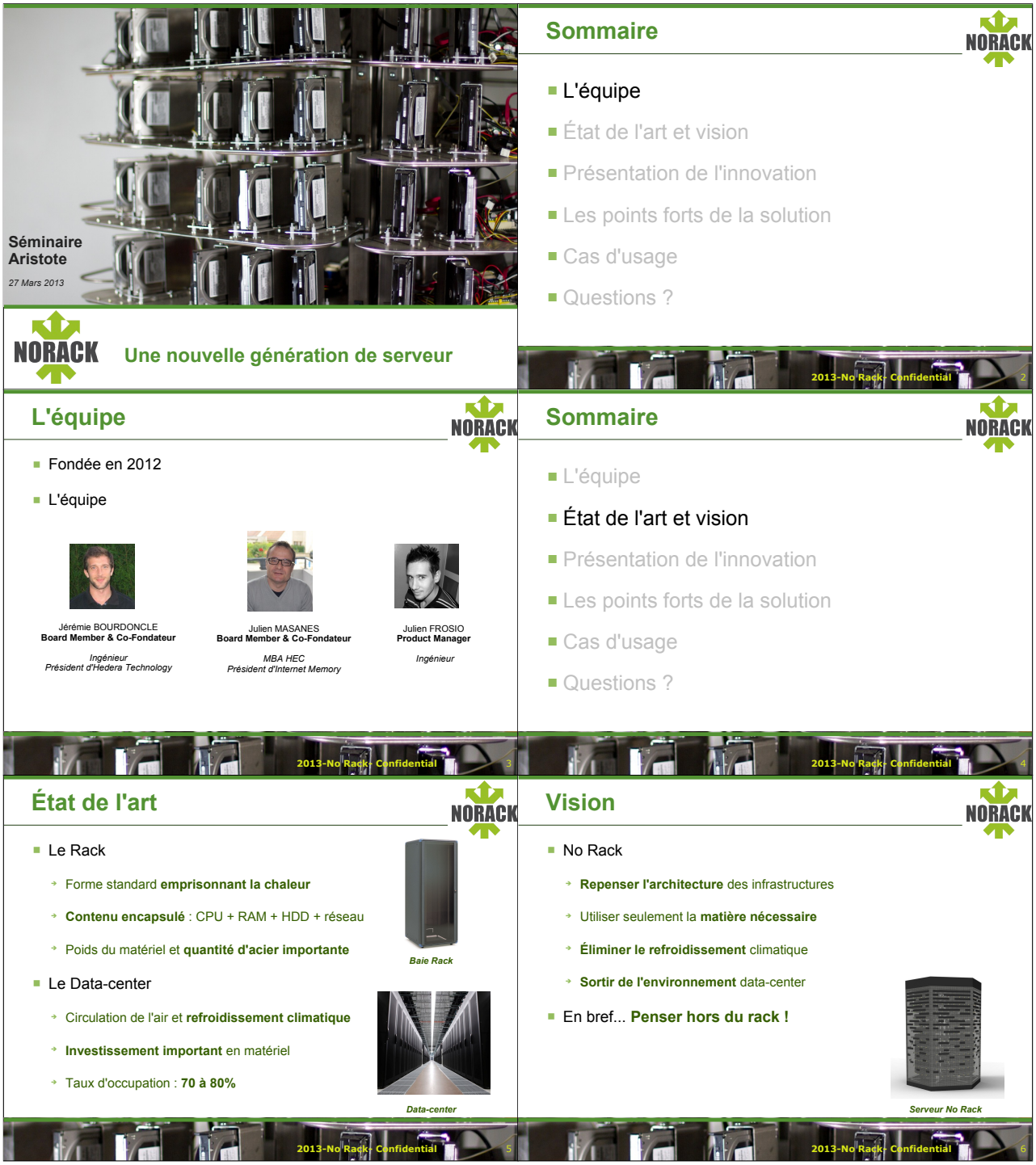
Next European dCache Workshop
27 May – 29 May
In Berlin

further reading
www.dCache.org

3.6 Jérémie Bourdoncle (NoRack)

Un système de stockage capacitif *green* et accessible

Jérémie Bourdoncle est le PDG de NoRack. Un segment du Big Data est encore trop peu exploré, celui du stockage des informations. Aujourd'hui, stocker des pétaoctects de données engendre des coûts relativement importants. Pour résoudre cette problématique, NoRack propose une innovation combinant stockage massif, basse consommation et *free cooling*.



Sommaire

- L'équipe
- État de l'art et vision
- Présentation de l'innovation
- Les points forts de la solution
- Cas d'usage
- Questions ?

Présentation de l'innovation 1/4

Prototype

Un châssis mécanique cylindrique

- Permet de faciliter la circulation de l'air et de minimiser la production de chaleur

Un système d'alimentation optimisé

- Permet de contrôler la consommation électrique du serveur

Une architecture en cluster

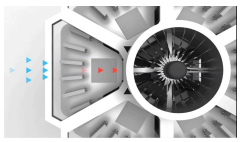
- Permet de fournir une configuration redondante
  - Basée sur 60 noeuds
  - **Gamme de stockage : 100 To à 1,1 Po**

Version finale

Présentation de l'innovation 2/4

- Le châssis mécanique
  - **Aucun refroidissement climatique** est nécessaire

Circulation de l'air au sein du serveur



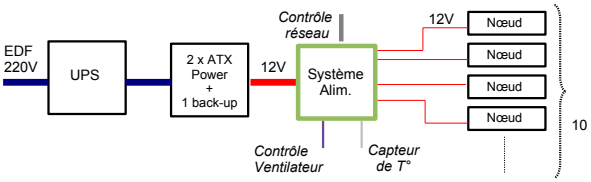
Effet venturi

Empreinte au sol : 1,13 m²

  - Peut fonctionner **hors d'un data-center**

Présentation de l'innovation 3/4

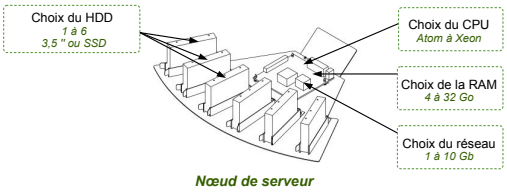
- Système d'alimentation optimisé
  - **Contrôle de chacun des nœuds** pour limiter la consommation
  - Utilisation d'alimentations ATX à leurs **rendements optimales**



Chaîne d'alimentation

Présentation de l'innovation 4/4

- Architecture en cluster
  - Permet d'avoir une **liberté de configuration** de chacun des nœuds



Nœud de serveur

  - **Accès direct aux équipements** en cas de maintenance

Sommaire

- L'équipe
- État de l'art et vision
- Présentation de l'innovation
- Les points forts de la solution
- Cas d'usage
- Questions ?

Les points forts de la solution

LE CHASSIS CYLINDRIQUE

- Facilite le passage de l'air entre les composants

LE SYSTEME D'ALIMENTATION

- Optimise les performances au niveau de la consommation d'énergie

L'ARCHITECTURE EN CLUSTER

- Permet l'utilisation de composants grand public

Consommation d'énergie indirecte	90%[1]	*Prévision de la consommation d'énergie liée au refroidissement*
Consommation d'énergie directe	85%[1]	*Réduction du besoin d'énergie pour le fonctionnement du serveur*
Coûts d'acquisition	20%[1]	*Limitation de l'utilisation de matériaux*
Coûts location d'un data-center	Non	*Option: Peut fonctionner hors d'un data-center*

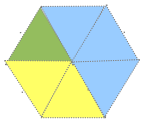
1 Estimation maximum des coûts

Sommaire

- L'équipe
- État de l'art et vision
- Présentation de l'innovation
- Les points forts de la solution
- Cas d'usage
- Questions ?

Cas d'usage 1/2

- Internet Memory

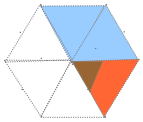


: **36 nœuds de stockage** : Atom CPU + 8 Go RAM + 5 x 3To HDD + Linux Debian + Hadoop, Hbase & Apache

: **15 nœuds de traitements de données** : Intel core i3 CPU + 16 Go RAM + 4 x 3To HDD + Linux Debian + logiciels propriétaires

: **20 nœuds de crawling** : Intel core i7 CPU + 16 Go RAM + 2 x 512 Go SSD + Linux Debian + logiciels propriétaires

- Hedera Technology



: **20 nœuds de stockage** : Atom CPU + 8 Go RAM + 5 x 3To HDD + Linux CentOS + OpenStack Ceph

: **7 nœuds de dev.** : Intel core i7 CPU + 32 Go RAM + 1 x 512 Go SSD + Linux CentOS + OpenStack Nova

: **3 nœuds de tests** : Intel core i3 CPU + 16 Go RAM + 1 x 256 Go SSD + 3 x 3To HDD + Linux CentOS

Cas d'usage 2/2

- Résultats :
  - En fonctionnement depuis **janvier 2012** (prototype up time : 99,2%)
  - Dans un environnement **hors data-center**
  - Consommation : **5 kW pour 750 To**
  - Environnement climatique (**opérationnel entre 10°C et 30°C**)

Questions ?

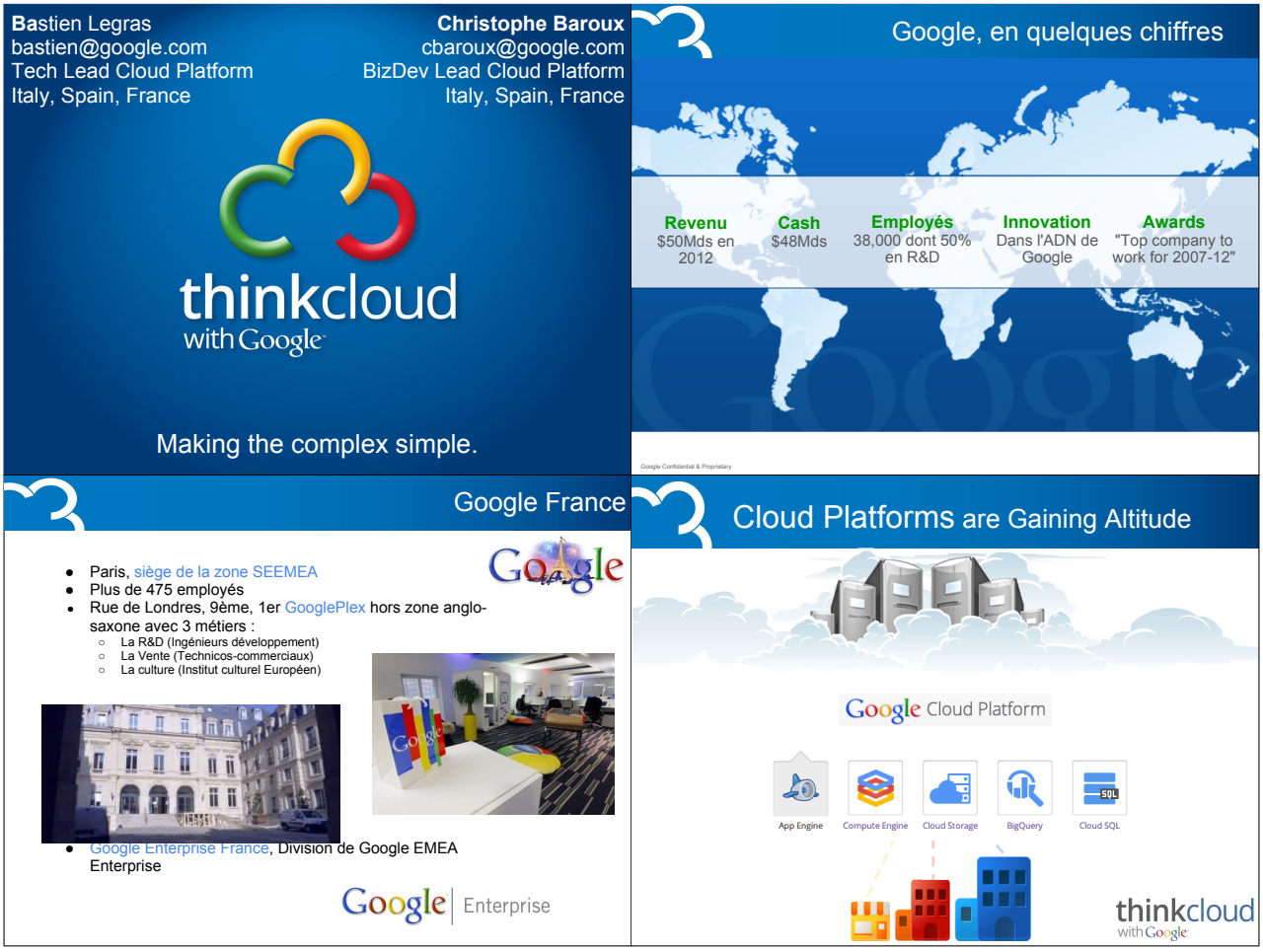
Contact : Jérémie BOURDONCLE
Email : jbourdoncle@no-rack.com

Web : www.no-rack.com
Twitter : @NoRack

3.7 Bastien Legras (Google)

BigQuery, le Big Data par Google

Bastien Legras est ingénieur solution responsable de l'offre *Cloud Platform* pour l'Europe du sud pour la division Google Enterprise. Avec une expérience de 5 ans dans le *cloud computing* (8 ans en IT), Bastien Legras est en charge d'accompagner les entreprises dans la construction de solutions *Cloud*. Google est une société incontournable dans les développements en tant que *Platform As A Service* (PaaS). Par son expérience dans le moteur de recherche et la continuité de service qui y est associée, le traitement du Big Data a été central dans le développement de Google. Des offres de plateformes autour de BigQuery et de Hadoop sont proposées par Google. A travers une illustration concrète, Google va présenter ses technologies et démontrer en quoi l'offre de service SAS de Google est pertinente dans le marché du Big Data.



Making The Complex Simple

Google Provides Full Library of APIs

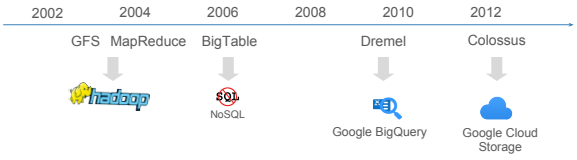
Crossing the Chasm

Google

Google confidential | Do not distribute

Google is Cloud

Google

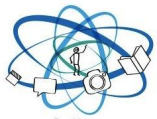


Google confidential | Do not distribute

Map-Reduce is 14 years old

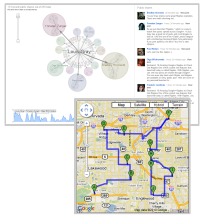
Google

Percolator
2010

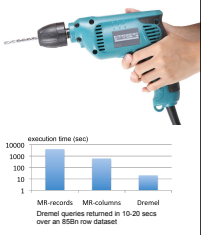


Caffeine
100 Pb Index
(circa 2010)

Pregel
June 2009



Dremel
2010



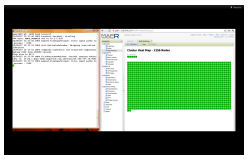
100x performance increase over MapReduce

[1] Large-scale Incremental Processing Using Distributed Transactions and Notifications, Daniel Peng and Frank Dabek, Google Inc, 2010
[2] Pregel: A System for Large-Scale Graph Processing, Grzegorz Malewicz, Matthew H. Austern, et al, Google Inc, 2010
[3] Dremel: Interactive Analysis of Web-Scale Datasets, Sergey Melnik and Andrey Gubarev et al., Google Inc, 2010

Google confidential | Do not distribute

The Hadoop Terasort Record

Google



MAPR TECHNOLOGIES

	Current	GCE
Servers	1460	1003
Disks	5840	4012
Cores	11680	1003
Time	1.02 mins	56 secs

1 http://www.eweek.com/c/a/Cloud-Computing/MapR-Integrates-Hadoop-Distro-with-Google-Compute-Engine-645801/
2 http://www.bloomberg.com/article/2012-10-24/aeJXnVPLO.HA.html

Google confidential | Do not distribute

The Results...

Google

Current Record

1460 1U servers x
$4K/server =

$5,840,000

Google MAPR TECHNOLOGIES

1003 *n1-standard-4-d* x
$.58/instance hour x
1 hour =

$582*

* actual pro-rata compute time < $10

Google confidential | Do not distribute

Thank You

code.google.com

thinkcloud
with Google


thinkcloud
with Google

3.8 Patrick Marques (HP)

Un cas concret chez nos clients : Hadoop

- structure Hadoop (comment ça fonctionne) ;
- caractérisation des *workloads* ;
- *sysing/architecture type* ;
- matériel : SL4500 – Moonshot.

Hadoop : Retour d'experience clients

Patrick Marques
Hyperscale Architecte avant-vente
27 Mars 2013

Hadoop – Architecture logicielle

What is Hadoop?

Apache Hadoop is an open source platform for data storage and processing that is…

- ✓ Scalable
- ✓ Fault tolerant
- ✓ Distributed

CORE HADOOP SYSTEM COMPONENTS

Hadoop Distributed File System (HDFS)
Self-Healing, High Bandwidth Clustered Storage

✚

MapReduce
Distributed Computing Framework

Has the Flexibility to Store and Mine Any Type of Data

- Ask questions across structured and unstructured data that were previously impossible to ask or solve
- Not bound by a single schema

Excels at Processing Complex Data

- Scale-out architecture divides workloads across multiple nodes
- Flexible file system eliminates ETL bottlenecks

Scales Economically

- Can be deployed on commodity hardware
- Open source platform guards against vendor lock

Hadoop

Son Ecosysteme

Hadoop est basé sur une technologie de Google :
Google File System (GFS) en 2003
MapReduce en 2004

OpenSource

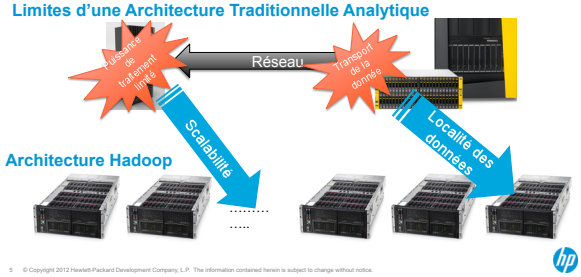
Hadoop est mantenant un écosysteme de produits OpenSource :
HDFS (Hadoop Filesysteme)
MapReduce : API Java
Pig
Hive
Hbase
Flume
Oozie
Sqoop
Zookeeper
Chuckwa
.....

Cloudera : CDH est une distribution supporté de produits Hadoop

---

Gains de l'architecture

Limites d'une Architecture Traditionnelle Analytique



Architecture Hadoop

Concept HADOOP 1/2

Point de vue Systeme

Export de données partagées

Les données sont stockés localement au serveur

2 types de serveurs Hadoop :
Serveur de métadonnées
Serveurs de données

Un fichier de données est "splitté" en blocks
64Mb ou 128Mb par block
Chaque block est répliqués (3 fois par défaut)

Solution de type cluster HPC, Cloud Computing : des centaines/milliers de serveurs banalisés

Concept HADOOP 2/2

Point de vue applicatif

Développement :

Les applications doivent être écrites dans un langage haut niveau
Abstraction de l'architecture materielle
Pas de communication explicite entre process

- approche différente de MPI

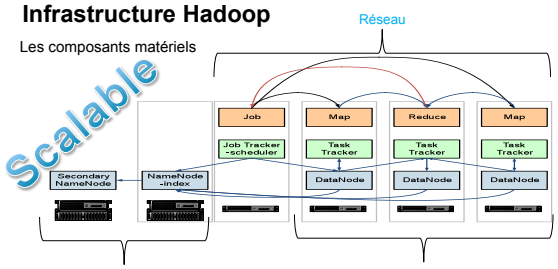
Production

Gestionnaire de batch

- le calcul doit s'exécuter de préférence sur le serveur qui possède la donnée a traiter
- Lors d'une panne d'un composant :
  - Le serveur doit être retiré de la production automatiquement
  - Il doit être réintégré automatiquement lors de son redémarrage
  - Si un échoue lié à cette panne, le jobdoit être ressoumis automatiquement sur un autre noeud

Infrastructure Hadoop

Les composants matériels



HDFS ?

HDFS est basé sur Google's GFS (<> Redhat GFS)
Se situe au dessus d'un filesysteme natif (ext4,xfs,…)
Ce n'est pas un systeme de fichier POSIX

Les Métadonnées sont stockés sur un serveur
Les Données sont eclatés sur plusieurs serveurs en block de 128 MB
⇒Performance

Chaque block est répliqué plusieurs fois (3 par défault) sur des serveurs différents
⇒Disponibilité

MapReduce

Abstraction via une API Java

- Le développeur doit écriture uniquement les fonctions Map et Reduce

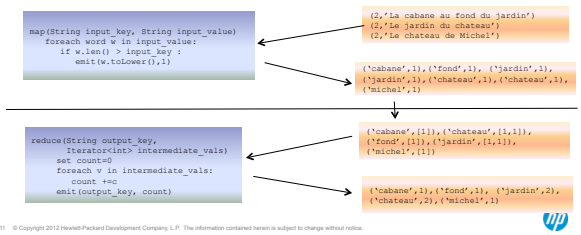
•Processus distribué de traitement de record qui s'execute en 2 phases : Map et Reduce

Gestionnaire de batch : JobTracker

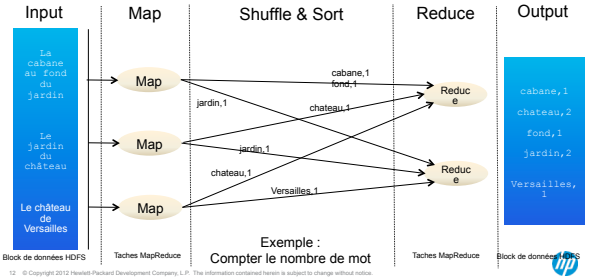
- Parallélisation automatique
- Continuité de service même en cas de défaillance de node
- Daemons :
  - JobTracker tourne sur le MasterNode
  - TaskTracker tourne sur tous les nodes

MAPREDUCE: EXEMPLE

Compter le nombre d'instance d'un même mot dans une phrase



MapRedUCE : TOPOLOGIE



Exploitation

Problématique de cluster

Une approche type HPC

Architecture TRADITIONNELLE

Solution high-end (BCS) :

- Serveurs Superdome
- Stockage XP
- Bases de donnée Oracle

Solution HADOOP

Solution Banalisée (ISS)

Des centaines/milliers de serveurs

Disque Interne

Couche Logicielle opensource

Transition identique au HPC dans les années 90
lors du passage des serveurs vectoriels (type Cray)
aux solutions de cluster de calcul.

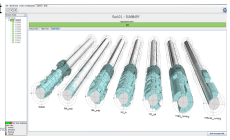
HP Insight Cluster Management Utility (CMU)

Hyperscale cluster lifecycle management software

Provision
- Simplified discovery, firmware audits
- Fast and scalable cloning

Monitor
- 'At a glance' view of entire system; zoom to component
- Customizable
- Lightweight and efficient

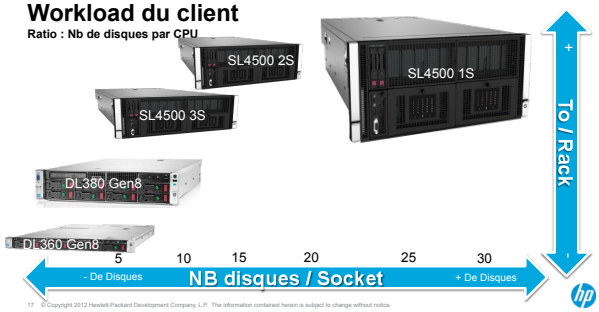
Control
- GUI and CLI options
- Easy, friction-less control of remote servers



- 10 years+ in deployment, included Top500 sites with1000's of nodes
- Built for Linux, with support for multiple Linux distributions
- HP supported, available as factory-integrated cluster option

Workload du client

Ratio : Nb de disques par CPU



Architecture Type Nœud Datanode

Retour expérience client

Serveur D	Config Standard
OS	Linux
CPU	Bi-Socket avec un **ratio** : 1 Disque dur pour 1 Core Intel
Disques	4 GB / Core
RAM	
Réseau	
Ctrl Disque	

2x E5-2440 (6C,2.4Ghz)
12x 3 To SAS 7.2K
48 GB : 6 x 8 GB

2x E5-2440 (6C,2.4Ghz)
15x 3 To SAS 7.2K
48 GB : 6 x 8 GB

Architecture Type Nœud Datanode

Retour expérience client

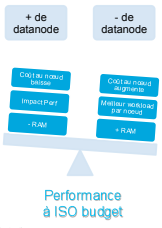
Serveur D	Config IO E
OS	Linux
CPU	Augmenter le nombre de disques par Core
Disques	4 GB / Core
RAM	
Réseau	
Ctrl Disque	

2x E5-2440 (6C,2.4Ghz)
25x 3 To SAS 7.2K
48 GB : 6 x 8 GB

Architecture Type Nœud Datanode

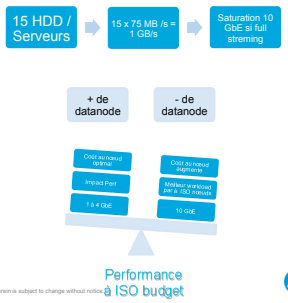
Impact de la RAM

Serveur D	Config Type
OS	4 GB / Core
CPU	
Disques	
RAM	
Réseau	
Ctrl Disque	



Architecture Réseau

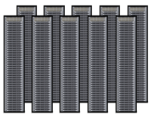
Figure2. UDA Performance Results



Breakthrough savings and simplicity

Radically changing economics of IT

Traditional x86
$3.3M



400 servers in 10 racks
20 switches
1,600 cables
91 kilowatts

89% less energy
94% less space
63% less cost
93% less HW complexity

Based on weighted average performance projections for workloads such as web serving, memcached, and Data Analytics. Cost estimates include infrastructure, space, and power and cooling costs over three years.

Example: "Gemini" Servers
$1.2M

1200 servers in 1 rack
2 switches
41 cables
9.9 kilowatts

BigData et Impact sur le marché

- Cost au GigaOctet optimal à une approche SAN
- Solutions basées sur une architecture équivalente
  - NAS Scalable
  - Storage object

Thank you

http://www.hp.com/go/hadoop

3.9 Peter Livaudais (ParStream)

ParStream, une base de données qui révolutionne la recherche en masse

Peter Livaudais est Directeur technique. L'allemand Parstream positionne son produit sur sa capacité de recherche rapide, sans comparaison à ce jour. Ce n'est pas du NoSql, ni du SQL complet mais les fonctionnalités évoluent pour en faire un des piliers du Big Data.

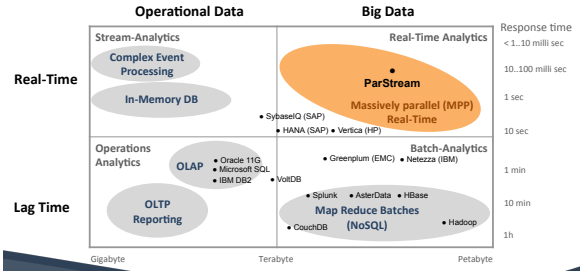
ParStream
REAL-TIME – LOW-LATENCY – HIGH THROUGHPUT

ParStream is THE Real-Time Big Data Analytics Platform.

We enable our customers to exploit Big Data opportunities through a unique parallel technology available on-premise, in the cloud and as a service

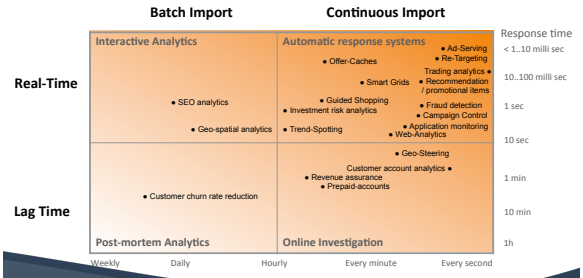
Analytics Landscape

Big Data analytics requires new technological solutions



PARSTREAM FOCUSES ON REAL-TIME ANALYTICS

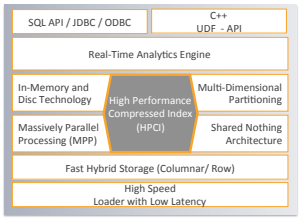
Real-Time Analytics Is Ultra-Fast on Continuously Imported Data



ARCHITECTURE BUILDING BLOCKS

ParStream is a Big Data Analytics Platform Based on a Unique High Performance Compressed Index

- Hybrid Columnar/Row Storage
- In Memory Technology
- Shared Nothing Architecture
- Standard Interfaces
- User Defined Functions
- **Unique High Performance Compressed Index**



HIGH PERFORMANCE COMPRESSED INDEX

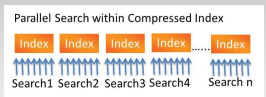
The Key to ParStream's Unmatched Performance

STANDARD DB INDEX ARCHITECTURE

- High Memory Requirements
- High Load on CPUs
- Time for Decompression

Not Suitable for Big Data Analytics

PARSTREAM INDEX ARCHITECTURE



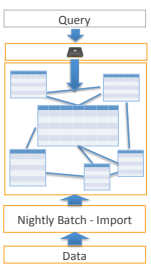
- + Low Memory Requirements
- + No Need for Decompression
- + Patent filing in process

Engineered for Big Data Analytics

PARALLEL ARCHITECTURE

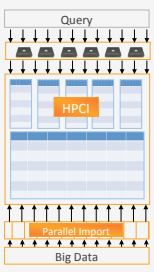
ParStream Overcomes Limitations of Traditional DW Architectures

STANDARD DW ARCHITECTURE



- − Long Query Runtime
- − Frequent Full Table Scans
- − Data is at Least 1 Day Old

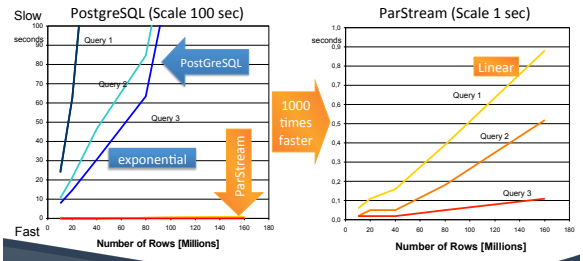
PARSTREAM ARCHITECTURE



- + Each Query Uses Multiple Processor Cores
- + Query execution using compressed indices
- + Continuous Import Assures Timeliness of Data

ORDERS OF MAGNITUDE FASTER

ParStream Outperforms PostgreSQL by a Factor of 1000
Delivering Results in Sub-Seconds on Large Data Volumes



ParStream

15

ONE LICENSE – FOUR WAYS TO DELIVER

Customer Choice – ParStream is a software-only product running on standard infrastructure. Together with partners ParStream is offered as an applicance and as a cloud service.

Price – ParStream's license price only depends on the maximum data volume a customer wants to store and analyze in a productive solution. Development, test and training instances are free of charge.

ParStream

CONFIDENTIAL 16

3.10 Sébastien Noury (Palantir)

Palantir Gotham, une plate-forme d'analyse issue de la Silicon Valley

Palantir aide les gouvernements et organisations commerciales à résoudre leurs problèmes d'analyse les plus complexes. Sa plate-forme Palantir Gotham est employée par des centaines d'organisations à travers le monde, dans des domaines tels que le renseignement, la prévention de fraude, la défense ou encore la réponse aux catastrophes naturelles. Palantir Gotham est un point d'entrée unique et transparent vers de multiples sources de données structurées, non structurées, temporelles ou géospatiales, qu'elles proviennent de flux temps réel, discontinus ou en quantité massive. Sebastien Noury, ingénieur déployé, présentera au travers d'un exemple concret comment Palantir Gotham permet aux analystes d'explorer ces données en temps réel, de découvrir et de consolider leurs connexions, ainsi que de collaborer de façon sécurisée. Résolument orientée Big Data, cette présentation se focalisera sur les capacités d'analyse offertes par l'exploration de données massives en temps réel.

Palantir

Palantir

Analyse en temps réel à l'échelle du petaoctet

Séminaire Aristote: "A la poursuite des Big Data"
27 Mars 2013, Ecole Polytechnique, France

Seb Noury – Forward Deployed Engineer – snoury@palantir.com

Palantir

Qui est Palantir?

Palantir

Qui est Palantir?

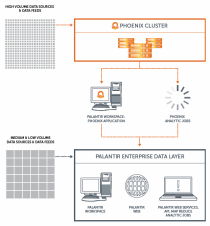
- Une poignée de passionnés venant de la Silicon Valley…
- Solvant les problèmes clés des plus grandes organisations
- … En réutilisant une plateforme technologique commune

- Croissance exponentielle, vaste majorité d'ingénieurs
- QG @ Palo Alto; Washington DC, New York, Los Angeles
- Londres, Canberra, Rome, Stockholm, Singapour, Berlin

Palantir

Notre approche

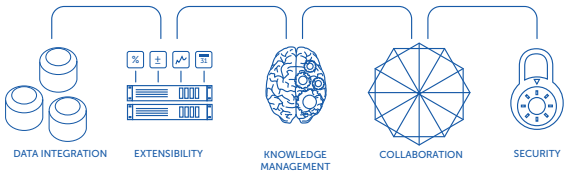
- Abolir les barrières traditionnelles
  - Défragmenter les sources de données
- Démocratiser les outils d'analyse
  - Simplifier la recherche et la collaboration
- Entrer en symbiose avec les experts
  - Facile à apprendre, rapide à maîtriser
- Résoudre les problèmes clés, efficacement



Palantir

Notre approche

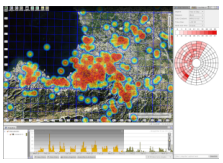
- Grandes institutions: problèmes critiques, clé = rapidité
- Plateformes Palantir: 90% de la solution déjà réalisée
- Equipes Palantir: intégration et itération jusqu'aux 100%



Palantir

Notre mission

- Permettre aux analystes d'interagir avec des Big Data
- Simplifier des processus d'analyse complexes
- Permettre une collaboration simple et sûre
- Propager le workflow des utilisateurs clés
- Travailler à l'échelle de grandes organisations



Palantir

Nos missions

Santé

Surveiller la prolifération atomique

Prévenir et limiter la propagation d'épidémies

Mesurer l'impact de catastrophes sur l'environnement

Palantir

Nos missions

Industrie, Finance

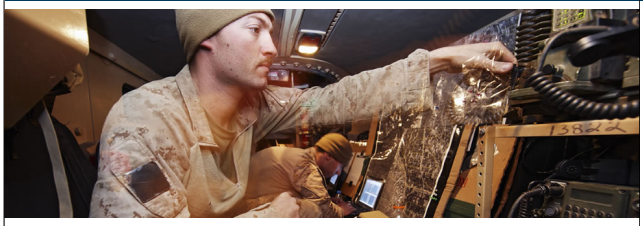
Détecter et prévenir les intrusions informatiques

Lutter contre la fraude et le blanchiment d'argent

Gérer les risques financiers, éviter une nouvelle bulle

Palantir

Nos missions



Armée, Forces de l'Ordre

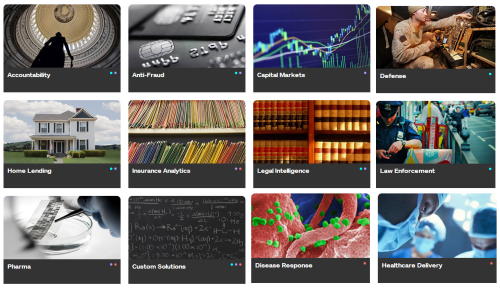
Lutter contre le crime organisé et la fraude fiscale

Démanteler les réseaux de trafic humain, de drogue, …

Supporter les troupes sur le front, partout dans le monde

Palantir

Nos missions



Palantir

Nos missions

Palantir Philanthropy + Team Rubicon

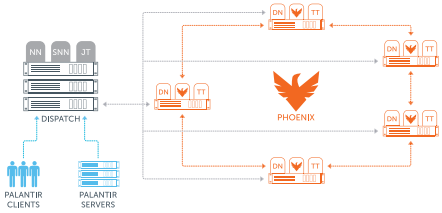
Organisation du support après l'ouragan Sandy

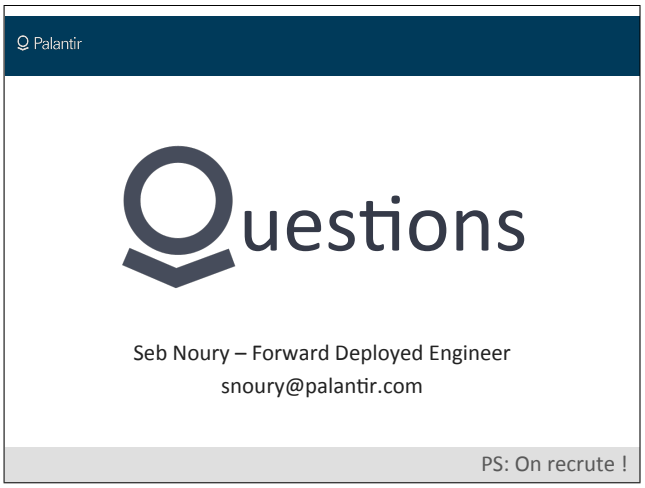
Palantir

Démo

Palantir Gotham + Phoenix

Analyse en temps réel à l'échelle du petaoctet




Palantir
Questions
Seb Noury – Forward Deployed Engineer
snoury@palantir.com
PS: On recrute !

3.11 Philippe Martin (Dell)

Peut-on faire passer des Big Data avec un modem 56kb/s

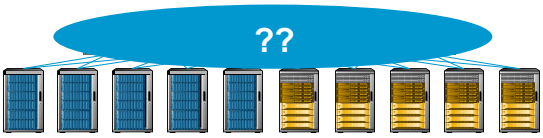
Philipe Martin est specialiste des ventes réseau. Big Data : le terme n'est pas immédiatement synonyme de réseau et ne focalise pas sur la capacité de traitement de ce dernier. Il convient néanmoins de constater que de grosses infrastructures de calcul et de traitement nécessitent une solution appropriée en terme de réseaux. Les besoins de débits, de latence, de performances de manière générale et de sécurisation d'architecture diffèrent en fonction des projets.

Dell est un acteur majeur de l'infrastructure, et compte toujours plus de clients pour ses solutions de mise en réseau ultra-performantes et sécurisées qui vous seront présentées ici, notamment à l'aide de réalisations d'architectures. Alors peut-on utiliser un modem 56kb/s pour son projet Big Data ?

Solutions Dell Networking pour le Big Data

Philippe MARTIN
Networking Sales Specialist - p_martin@dell.com

Peut-on faire passer des big data avec un modem 56kbs ?!



- Le réseau est souvent l'oublié d'un projet Big Data
  - Priorité à l'achat de nœuds/serveurs
  - Quid des performances réseau ?
  - Quid de l'infrastructure en cas de défaillance ?
  - Comment assurer le management réseau du Big Data ?

Dell Networking : aperçu rapide
- Notre présence dans les réseaux
  - Historique et positionnement
  - Stratégie

Notre vision réseaux du Big Data
- Active Fabric
  - Architecture type
  - Solutions
  - Nos points forts

Perspectives
- Automation
- Notre vision pour demain

Notre présence dans les réseaux

- Etre un acteur majeur de l'infrastructure
  - Serveurs
  - Stockage
  - **Réseau**
- **Plus de 10 ans d'expérience**
  - 4ème constructeur mondial de switches
  - 3ème constructeur mondial de switches 10Gb
  - 3ème constructeur mondial de switches 40Gb
  - Présence au Magic Quadrant Gartner pour le Data Center Networking
- Nous souhaitons représenter une **alternative crédible, qualitative et cohérente** en matière de réseaux

Nous avons un point de vue unique

Virtual network architecture

- Disposer d'un catalogue complet de solutions LAN
  - Campus
  - Data Center / calcul
- Proposer une vision **simple, ouverte, ultra-performante et ultra-sécurisée** des réseaux : Active Fabric

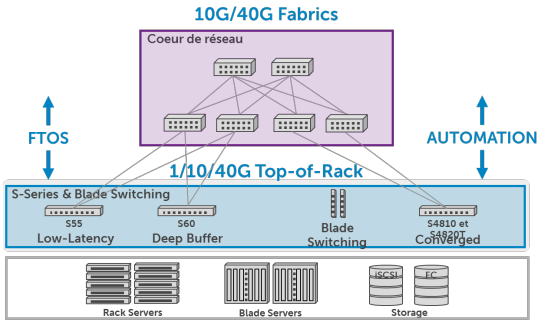
Agenda

Dell Networking : aperçu rapide
- Notre présence dans les réseaux
  - Historique et positionnement
  - Stratégie

Notre vision réseaux du Big Data
- Active Fabric
  - Architecture type
  - Solutions
  - Nos points forts

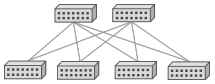
Perspectives
- Automation
- Notre vision pour demain

Notre vision du réseau Big Data : Activ Fabric



Big Data : Cœur de réseau distribué

- Les avantages du cœur de réseau distribué :
  - Coût
  - Evolutivité
  - Résilience



- **Pour une infrastructure Big Data simple, ulta-performante et ultra sécurisée**
  - Z9000
    - 32x 40Gb (jusqu'à 128x 10Gb)
    - L2/L3 IPv4/IPv6
    - VLT
      - Redondance & liens actifs-actifs L2/L3
      - Continuité de services 24x7

  - S4810
    - 48x 1Gb/10Gb + 4x40Gb
    - L2/L3 IPv4/IPv6
    - VLT

Big Data : switches Top-of-Racks 1/10/40Gb

- Ultra-Low Latency 10Gb et VLT : S4810
  - Uplinks 10Gb/40Gb
  - Redondance Actif-Actif des liens vers les nœuds et le cœur (VLT)

- Forte concentration 10Gb cuivre et VLT : S4820T
  - 48x 1Gb/10Gb Cuivre + 4x40Gb
  - L2/L3 IPv4/IPv6
  - VLT
    - Liens actifs-actifs L2/L3
    - Continuité de services 24x7

- Evolutivité 24 à 64 ports 10Gb : 8100
  - Évolutivité par ajout de module
  - Stackable

Big Data : switches Top-of-Racks 1Gb/10Gb

- Low Latency : S55
  - 48 ports 1Gb
  - 2 emplacements modulaires pour stack et/ou uplinks 10Gb
  - Stackable

- Deep-buffers : S60
  - 48 ports 1Gb
  - 2 emplacements modulaires pour stack et/ou uplinks 10Gb
  - stackable

- Pour les configurations plus classiques : 7000
  - 24 ou 48 ports 1Gb
  - 2 emplacements modulaires pour stack et/ou uplinks 10Gb
  - Stackable

Big Data : Management

- Active Fabric Manager
  - Point unique de paramétrage de la fabrique
  - Provisionning automatisé des configurations
  - Design de Templates Acitv Fabric
- OMNM
  - Cartographie des équipements
    - Inventaire et historique des niveaux de version et fichiers de configuration
  - Gestion de la configuration des équipements
  - Remontée des statistiques
    - Sflow
    - NetFlow
    - Génération d'alertes

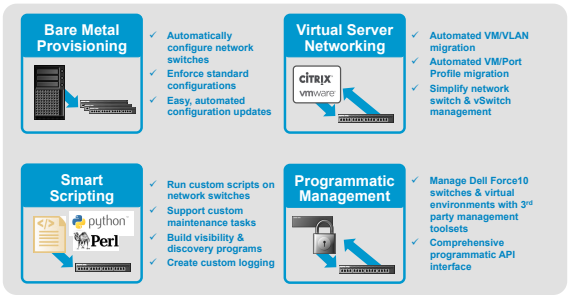
Agenda

Dell Networking : aperçu rapide
- Notre présence dans les réseaux
  - Historique et positionnement
  - Stratégie

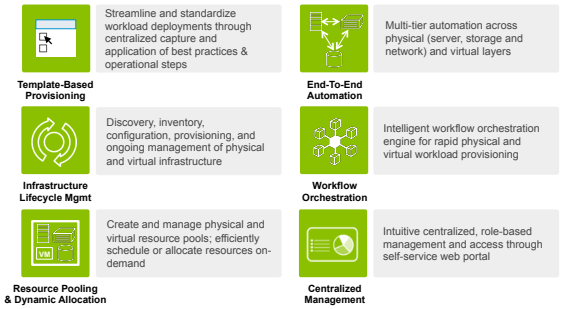
Notre vision réseaux du Big Data
- Active Fabric
  - Architecture type
  - Solutions
  - Nos points forts

Perspectives
- Automation
- Notre vision pour demain

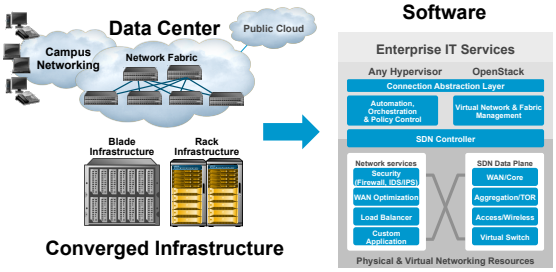
Perspectives actuelles : apporter de la valeur à son réseau Big Data - Automatisation



Orchestration : Active System Manager



Allocation dynamique de ressources Big Data par projet : SDN et OpenFlow



Thank you

Philippe MARTIN – p_martin@dell.com

ARISTOTE Association Loi de 1901. Siège social : CEA-DSI CEN Saclay Bât. 474, 91191 Gif-sur-Yvette Cedex.
Secrétariat : Aristote, École Polytechnique, 91128 Palaiseau Cedex.
Tél. : +33(0)1 69 33 99 66 Fax : +33(0)1 69 33 99 67 Courriel : Marie.Tetard@polytechnique.edu
Site internet http://www.association-aristote.fr