Proposition de recherche doctorale

Auto-organisation dynamique de services dans un environnement de Cloud à destination de flux de données personnelles

Mots clés :

- **Directeur de thèse** : Eric Gressier-Soudan
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'Informatique, Signal et Image, Electronique et Télécommunication
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Actuellement, de plus en plus de personnes génèrent un nombre croissant de flux de données à partir de terminaux informatiques personnels et portatifs (smartphones et tablettes). Ces données sont le plus souvent destinées à un usage précis (par exemple envoyer un message et/ou signaler une position géographique). Néanmoins, de même que la fouille de données industrielle a d'abord consisté à valoriser des données relevées pour un autre but immédiat (la facturation de consommation, les mesures de production ou le contrôle de processus industriels) grâce à des algorithmes d'apprentissage statistique, une telle évolution est prévisible pour les données personnelles. En fait, elle a souvent déjà lieu plus ou moins à l'insu des utilisateurs, pour assurer la rentabilité économique des services gratuits qu'ils consomment. Cependant, on peut considérer que l'exploitation des données personnelles pourrait être encore développée si elle restait sous le contrôle de la personne concernée elle-même. Poussée à l'extrême, la pratique du lifelogging consiste à enregistrer le maximum d'informations en flux continu. L'exploitation de ce genre de données amène des besoins en stockage et traitement de données hors de portée des dispositifs qui génèrent ces données. De plus, l'internet des objets sera lui-aussi la source d'un nombre croissant de données qui pourront désormais être stockées sur le réseau. L'utilisateur concerné n'ayant pas forcément les compétences techniques et/ou l'envie de gérer l'infrastructure nécessaire, il/elle externalisera la gestion de celle-ci sur une architecture mutualisée de type Cloud [bell]. Si l'externalisation des tâches de stockage et surtout de calcul est généralement intéressante financièrement grâce aux économies d'échelle et à la mutualisation, elle pose des problèmes d'optimisation et de visibilité des coûts. Pour une utilisation personnelle, les tâches de fouille de données personnelles sont optionnelles et l'on voudra pouvoir en estimer les coûts a priori pour choisir d'effectuer ou non celles-ci. De plus l'utilisateur n'ayant pas forcément de compétence dans ce domaine, l'optimisation de l'exécution doit elle-même être réalisée de façon complètement automatique. Ces nouveaux usages (lifelogging et internet des objets) sont donc en train de susciter des demandes en matière de fouille de données "démocratisée" c'est-à-dire mise à la disposition de personnes qui n'ont pas l'expertise technique sur les enjeux du déploiement. [bell] Gordon Bell, entretien à Scientific America 2011-05-04.

Résumé du projet de recherche (Langue 2)

Jusqu'à maintenant, les données disponibles pour étudier et optimiser finement les usages de ressources sur le cloud n'étaient pa disponibles pour les utilisateurs. En effet, les fournisseurs d'infrastructure ne mettaient à disposition que le minimum nécessaire pour la facturation. Cependant, le marché des technologies Cloud est arrivé à la maturité suffisante pour voir émerger des solution libres (i.e. OpenStack) qui mettent entre les mains des utilisateurs toutes les informations que ceux-ci peuvent désirer [ceilometer]. C'est de cette disponibilité nouvelle que nous entendons tirer partie. Le passage à un mode de facturation à l'usage des ressources informatiques amène la question de la prévisibilité des coûts. Il s'agit d'un domaine récent dans lequel des start-ups se sont engagées en 2012. Pour l'instant, les services proposés [cloudyn][newvem] sont très basiques et se contentent d'indiquer les utilisations en fonction d'heures de la journée et des jours de la semaine. Ils ne permettent pas de réagir dynamiquement en fonction de tendance localisées dans le temps. L'approche que nous proposons d'étudier au cours de cette thèse consiste à considérer les services qui composent le système hébergé comme des agents qui génèrent chacun des données (les mesures de consommation de ressources cloud). Le système composé d'un ensemble de services sera considéré en tant que tel et l'on apprendra les corrélations entre ceux-ci afin d'obtenir de meilleurs résultats en prévision que si l'on considérait chaque service indépendamment des autres. En effet, les conditions extérieures qui affectent les agents (ici les services) sont généralement corrélées entre elles d'une façon qu'il est possible d'apprendre au fil de l'eau[Tatsu2011]. Des algorithmes coûteux en temps de calcul peuvent être utilisés offline pour indexer les données collectées au fil de l'eau. Ceci permet aux algorithmes de recherche ayant des contraintes en temps d'exécution d'être plus performants selon une approche classique du Big Data. 1. Dans un premier temps, on se limitera à considérer les services comme des capteurs passif des grandeurs mesurées. On effectuera alors des tâches de prévision [caron] pour donner à l'utilisateur le moyen de décider ou non d'effectuer des tâche de data-mining sur ses lifelogs et selon un plan d'exécution adapté. 2. Dans un deuxième temps, on considèrera les services comme des agents pouvant se déplacer/dupliquer et sur lesquels il est possible d'appliquer un contrôle, comme des cyber-physical systems. La localisation, plus ou moins proche de données des code pouvant être un aspect important du coût d'exécution (en temps -latence- et en débit facturé). Dans le cas, le plus fréquent, de serveurs virtualisés, les comportements des autres serveurs virtuels partageant la même machine physique auront aussi une influence sur les ressources disponibles (à coût facturé constant !). On tiendra aussi compte de ce phénomène lors des prévisions et pour déterminer les politiques de transfert de charge. 3. Avec la plus grande complexité des architectures orientées service, la résilience des systèmes dépendants de nombreux autres services peut devenir problématique. C'est le même genre de problématique que pour les cyber-physical systems [cps-dc] : le résilience du système complexe ne peut pas se baser sur celles de ces composants aussi fiables soient-ils individuellement. La documentation d'Hystrix [hystrix] donne l'exemple suivant : un système dépendance de 30 services ayant chacun une disponibilité de 99.99% n'aura (dans le cas de décorrélation des sources d'erreur !) qu'une disponibilité de 99.99%^{30}=99.7% soit deux heures d'indisponibilité par mois. Là encore, une étude au fil de l'eau des conditions d'indisponibilité devrait pouvoir permettre d'optimiser la résilience du système. [caron] Forecasting for Cloud computing on-demand resources based on pattern matching E.Caron et al INRIA Technical Report 2010 [ceilometer] https://wiki.openstack.org/wiki/Ceilometer [cps-dc] Cyber Physical Systems: Design Challenges, Edward A. Lee, ISORC, 2008 [hystrix] https://github.com/Netflix/Hystrix/wiki utilisé par Netflix [Tatsu2011] spatio-temporal analysis and modeling of short-term wind power forecast errors J.Tastu et al. Wind Energy Volume 14, Issue 1, 2011

Informations complémentaires (Langue 1)

C'est envisagé, mais pas encore mis en place formellement.

Informations complémentaires (Langue 2)

La thèse correspond à une coopération entre deux laboratoires le LISITE et le CEDRIC. Elle sera encadrée par Mr Eric Gressier Soudan, professeur HDR (Cedric, CNAM) et par Mr Bernard Hugueney et Mme Raja Chiky, enseignants chercheurs au LISITE, ISEP.