

Conception et mise en place d'un système Workflow pour l'environnement cloud

Institution



Zerrouki Djamel

20 mars 2019

Abstract

Abstract goes here

Dedication

To mum and dad

Declaration

I declare that..

Acknowledgements

I want to thank...

Table des matières

1	Concepts fondamentaux de cloud et de workflow	7
1.1	Introduction	7
1.2	cloud computing	7
1.2.1	Concept du cloud computing	7
1.2.1.1	Vers une définition du cloud computing	8
1.2.1.2	Caractéristiques principales du cloud computing	8
1.2.2	Technologies connexes	10
1.2.3	Modèles du cloud computing	10
1.2.3.1	Modèles de service du cloudcomputing	10
1.2.3.2	Modèles de déploiement	11
1.3	Workflow et systèmes de gestion de Workflow	12
1.3.1	Concepts de base et définitions de Workflow	12
1.3.1.1	Définitions de workflow	13
1.3.1.2	Les types de workflow	13
2	Modélisations et Réseau de Petri	15
3	Implémentation	16

Chapitre 1

Concepts fondamentaux de cloud et de workflow

1.1 Introduction

Le cloud computing, traduit le plus souvent en français par "informatique dans les nuages", "informatique dématérialisée" ou encore "infonuagique", est un domaine qui regroupe un ensemble de techniques et de pratiques consistant à accéder, en libre-service, à du matériel ou à des logiciels informatiques, à travers une infrastructure réseau (Internet). Ce concept rend possible la distribution des ressources informatiques sous forme de services pour lesquels l'utilisateur paie uniquement pour ce qu'il utilise. Ces services peuvent être utilisés pour exécuter des applications scientifiques et commerciales, souvent modélisées sous forme de workflows. (see GOOSSENS et al. 1993, p10)

Ce chapitre présente une introduction au cloud computing et au workflow, nécessaire pour la compréhension générale de ce rapport.

Tout d'abord, nous présentons dans la section 1.2 une introduction au paradigme du cloud computing. Nous donnons un aperçu général du cloud computing, y compris sa définition, ses caractéristiques principales et une comparaison avec les technologies connexes. Nous présentons les différents modèles de service, les différents modèles de déploiement, ainsi que les différents acteurs du cloud computing. Nous résumons quelques challenges de recherche en cloud computing. Par la suite, nous présentons, dans la section 1.3, une introduction au workflow et systèmes de gestion de workflow. Nous donnons le concept du workflow, sa définition, et l'architecture de référence d'un système de gestion de workflows. Nous énumérons quelques systèmes de gestion de workflows existant dans les grilles et clouds et, finalement, nous résumons l'intérêt du cloud pour les workflows.

1.2 cloud computing

1.2.1 Concept du cloud computing

L'idée principale du cloud est apparue dans les années 60, où le professeur John McCarthy avait imaginé que les ressources informatiques seront fournies comme des services d'utilité publique (Garfinkel, 1999). C'est ensuite, vers la fin des années 90, que ce concept a pris de l'importance avec l'avènement du grid computing

(Foster, 1999). Le terme cloud est une métaphore exprimant la similarité avec le réseau électrique, dans lequel l'électricité est produite dans de grandes centrales, puis disséminée à travers un réseau jusqu'aux utilisateurs finaux. Ici, les grandes centrales sont les Datacenter, le réseau est le plus souvent celui d'Internet et l'électricité correspond aux ressources informatiques. Le cloud computing n'est véritablement apparu qu'au cours de l'année 2006 (Vouk, 2008) avec l'apparition d'Amazon EC2 (Elastic Compute cloud). C'est en 2009 que la réelle explosion du cloud survint avec l'arrivée sur le marché de sociétés comme Google (Google App Engine), Microsoft (Microsoft Azure), IBM (IBM Smart Business Service), Sun (Sun cloud) et Canonical Ltd (Ubuntu Enterprise cloud). D'après une étude menée par Forrester (Ried, 2011), le marché du cloud computing s'élevait à environ 5,5 milliards de dollars en 2008, il devrait atteindre plus de 150 milliards d'ici 2020, comme l'illustre la figure 1.1.

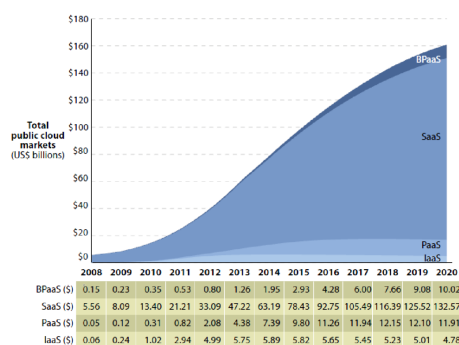


FIGURE 1.1 – Prévisions de la taille du marché du cloud computing public (Ried, 2011).

1.2.1.1 Vers une définition du cloud computing

Beaucoup de chercheurs ont tenté de définir le cloud computing (Geelan, 2008 ; McFedries, 2008 ; Buyya, 2009 ; Armbrust, 2010). La plupart des définitions attribuées à ce concept semblent se concentrer seulement sur certains aspects technologiques. L'absence d'une définition standard a généré non seulement des exagérations du marché, mais aussi des confusions. Pour cette raison, il y a eu récemment des travaux sur la normalisation de la définition du cloud computing, à l'exemple de Vaquero et coll (Vaquero, 2009) qui ont comparé plus de 20 définitions différentes et ont proposé une définition globale. En guise de synthèse des différentes propositions données dans la littérature, nous introduisons une définition mixte, qui correspond aux différents types de cloud considérés dans les travaux réalisés dans cette thèse.

Nous définissons le cloud comme un modèle informatique qui permet d'accéder, d'une façon transparente et à la demande, à un pool de ressources hétérogènes physiques ou virtualisées (serveurs, stockage, applications et services) à travers le réseau. Ces ressources sont délivrées sous forme de services reconfigurables et élastiques, à base d'un modèle de paiement à l'usage, dont les garanties sont offertes par le fournisseur via des contrats de niveau de service (SLA, Service Level Agreement).

1.2.1.2 Caractéristiques principales du cloud computing

Le cloud computing possède les caractéristiques suivantes :

- **Accès en libre-service à la demande.** Le cloud computing offre des ressources et services aux utilisateurs à la demande. Les services sont fournis de façon automatique, sans nécessiter d'interaction humaine (Mell, 2011).
- **Accès réseau universel.** Les services de cloud computing sont facilement accessibles au travers du réseau, par le biais de mécanismes standard, qui permettent une utilisation depuis de multiples types de terminaux (par exemple, les ordinateur portables, tablettes, smartphones) (Mell, 2011).
- **Mutualisation de ressources** (Pooling). Les ressources du cloud peuvent être regroupées pour servir des utilisateurs multiples, pour lesquels des ressources physiques et virtuelles sont automatiquement attribuées (Mell, 2011). En général, les utilisateurs n'ont aucun contrôle ou connaissance sur l'emplacement exact des ressources fournies. Toutefois, ils peuvent imposer de spécifier l'emplacement à un niveau d'abstraction plus haut.
- **Scalabilité et élasticité.** Des ressources supplémentaires peuvent être automatiquement mises à disposition des utilisateurs en cas d'accroissement de la demande (en réponse à l'augmentation des charges des applications) (Geelan, 2008), et peuvent être libérées lorsqu'elles ne sont plus nécessaires. L'utilisateur a l'illusion d'avoir accès à des ressources illimitées à n'importe quel moment, bien que le fournisseur en définisse généralement un seuil (par exemple : 20 instances par zone est le maximum possible pour Amazon EC2).
- **Autonome.** Le cloud computing est un système autonome et géré de façon transparente pour les utilisateurs. Le matériel, le logiciel et les données au sein du cloud peuvent être automatiquement reconfigurés, orchestrés et consolidés en une seule image qui sera fournie à l'utilisateur (Wang, 2008).
- **Paiement à l'usage.** La consommation des ressources dans le cloud s'adapte au plus près aux besoins de l'utilisateur. Le fournisseur est capable de mesurer de façon précise la consommation (en durée et en quantité) des différents services (CPU, stockage, bande passante, . . .) ; cela lui permettra de facturer l'utilisateur selon sa réelle consommation (Armbrust, 2009).
- **Fiabilité et tolérance aux pannes.** Les environnements cloud tirent parti de la redondance intégrée du grand nombre de serveurs qui les composent en permettant des niveaux élevés de disponibilité et de fiabilité pour les applications qui peuvent en bénéficier (Buyya, 2008).
- **Garantie QoS.** Les environnements de cloud peuvent garantir la qualité de service pour les utilisateurs, par exemple, la performance du matériel, comme la bande passante du processeur et la taille de la mémoire (Wang, 2008).
- **Basé-SLA.** Les clouds sont gérés dynamiquement en fonction des contrats d'accord de niveau de service (SLA) (Buyya, 2008) entre le fournisseur et l'utilisateur. Le SLA définit des politiques, telles que les paramètres de livraison, les niveaux de disponibilité, la maintenabilité, la performance, l'exploitation, ou autres attributs du service, comme la facturation, et même des sanctions en cas de violation du contrat. Le SLA permet de rassurer les utilisateurs dans leur idée de déplacer leurs activités vers le cloud, en fournissant des garanties de QoS. Après avoir présenté les caractéristiques essentielles d'un service cloud, nous présentons, brièvement, dans la section suivante, quelques technologies connexes aux clouds.

1.2.2 Technologies connexes

1.2.3 Modèles du cloud computing

1.2.3.1 Modèles de service du cloud computing

XaaS (X as a Service) représente la base du paradigme du cloud computing, où X représente un service tel qu'un logiciel, une plateforme, une infrastructure, un Business Process, etc. Nous présentons, dans cette section, quatre modèles de services (Rimal, 2009), à savoir : (1) Logiciel en tant que services SaaS (Software as a Service), où le matériel, l'hébergement, le framework d'application et le logiciel sont dématérialisés, (2) Plateforme en tant que service PaaS (Platform as a Service), où le matériel, l'hébergement et le framework d'application sont dématérialisés, (3) Infrastructure en tant que service IaaS (Infrastructure as a Service) et (4) Matériel en tant que service HaaS (Hardware as a Service), où seul le matériel (serveurs) est dématérialisé dans ces deux derniers cas. La figure 1.2 montre le modèle classique et les différents modèles de service de cloud

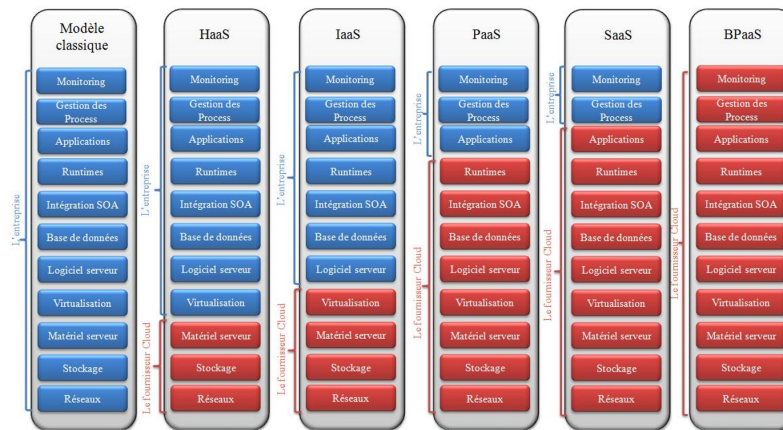


FIGURE 1.2 – Les services XaaS du cloud computing

1. Software as a Service (SaaS) :

Ce modèle de service est caractérisé par l'utilisation d'une application partagée qui fonctionne sur une infrastructure Cloud. L'utilisateur accède à l'application par le réseau au travers de divers types de terminaux (souvent via un navigateur web). L'administrateur de l'application ne gère pas et ne contrôle pas l'infrastructure sous-jacente (réseaux, serveurs, applications, stockage). Il ne contrôle pas les fonctions de l'application à l'exception d'un paramétrage de quelques fonctions utilisateurs limitées. On prend comme exemple les logiciels de messagerie au travers d'un navigateur comme Gmail ou Yahoo mail.

2. Platform as a Service (PaaS) :

L'utilisateur a la possibilité de créer et de déployer sur une infrastructure Cloud PaaS ses propres applications en utilisant les langages et les outils du fournisseur. L'utilisateur ne gère pas ou ne contrôle pas l'infrastructure Cloud sous-jacente (réseaux, serveurs, stockage) mais l'utilisateur contrôle l'application déployée et sa configuration. Comme exemple de PaaS, on peut citer un des plus anciens -IntuitQuickbase- qui permet de déployer ses applications

bases de données en ligne ou -Google Apps Engine (GAE)- pour déployer des services Web.

Dans ces deux cas l'utilisateur de ces services n'a pas à gérer des serveurs ou des systèmes pour déployer ses applications en ligne et dimensionner des ressources adaptées au trafic.

3. Infrastructure as a Service (IaaS) :

L'utilisateur loue des moyens de calcul et de stockage, des capacités réseau et d'autres ressources indispensables (partage de charge, pare-feu, cache). L'utilisateur a la possibilité de déployer n'importe quel type de logiciel incluant les systèmes d'exploitation. L'utilisateur ne gère pas ou ne contrôle pas l'infrastructure Cloud sous-jacente mais il a le contrôle sur les systèmes d'exploitation, le stockage et les applications. Il peut aussi choisir les caractéristiques principales des équipements réseau comme le partage de charge, les pare-feu, etc. L'exemple emblématique de ce type de service est Amazon Web Services qui fournit du calcul (EC2), du stockage (S3, EBS), des bases de données en ligne (SimpleDB) et quantité d'autres services de base. Il est maintenant imité par de très nombreux fournisseurs.

4. Points forts et Points faibles des services cloud :

	Points forts	Points faibles
SaaS	Pas d'installation Plus de licence	Logiciel limité Sécurité Dépendance de prestataire
PaaS	Pas d'infrastructure Nécessaire Pas d'infrastructure Nécessaire	Limitation des langages Pas de personnalisation dans la configuration des machines virtuelles
IaaS	Administration Personnalisation Flexibilité d'utilisation	Sécurité Besoin d'un administrateur système

TABLE 1.1 – Points forts et Points faibles des services Cloud

1.2.3.2 Modèles de déploiement

Selon la définition du cloud computing donnée par le NIST (Mell, 2011), il existe quatre modèles de déploiement des services de cloud, à savoir : cloud privé, cloud communautaire, cloud public et cloud hybride, comme illustré dans la figure 1.3.

1. Cloud privé :

L'ensemble des ressources d'un cloud privé est exclusivement mis à disposition d'une entreprise ou organisation unique. Le cloud privé peut être géré par l'entreprise elle-même (cloud privé interne) ou par une tierce partie (cloud privé externe). Les ressources d'un cloud privé se trouvent généralement dans les locaux de l'entreprise ou bien chez un fournisseur de services. Dans ce dernier cas, l'infrastructure est entièrement dédiée à l'entreprise et y est accessible via un réseau sécurisé (de type VPN). L'utilisation d'un cloud privé permet de garantir, par exemple, que les ressources matérielles allouées ne seront jamais partagées par deux clients différents.

2. Cloud communautaire :

L'infrastructure d'un cloud communautaire est partagée par plusieurs organisations indépendantes ayant des intérêts communs. L'infrastructure peut être gérée par les organisations membres ou par un tiers. L'infrastructure peut être située, soit au sein des dites organisations, soit chez un fournisseur de services.

3. Cloud public :

L'infrastructure d'un cloud public est accessible à un large public et appartient à un fournisseur de services. Ce dernier facture les utilisateurs selon la consommation et garantit la disponibilité des services via des contrats SLA.

4. Cloud hybride :

L'infrastructure d'un cloud hybride est une composition de plusieurs clouds (privé, communautaire ou public). Les différents clouds composant l'infrastructure restent des entités uniques, mais sont reliés par une technologie standard ou propriétaire permettant ainsi la portabilité des données ou des applications déployées sur les différents clouds.

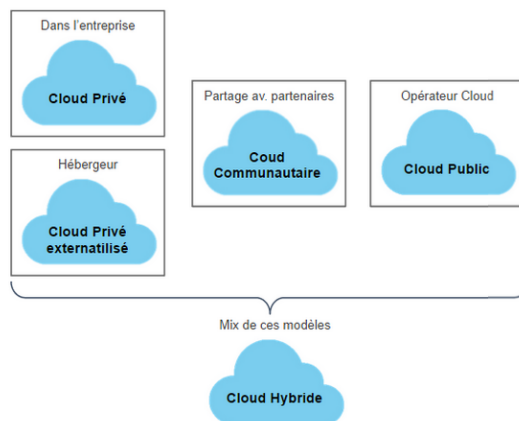


FIGURE 1.3 – Modèles de déploiement du cloud computing

1.3 Workflow et systèmes de gestion de Workflow

1.3.1 Concepts de base et définitions de Workflow

La notion de workflow (traduit en français par "flux de travail") est apparue dans l'industrie de l'image électronique et de la gestion de production assistée par ordinateur (GW, 1998). Ce concept a donc été créé dans le but d'automatiser les procédures de travail au sein des organisations. L'idée d'enchaîner différentes tâches pour réaliser un traitement complexe est pertinente. De plus, dans les infrastructures actuelles distribuées, gérant des ressources hétérogènes, telles que le cloud computing, bénéficier d'un environnement autorisant la définition et l'exécution des chaînes de traitement constitue une des fonctionnalités essentielles recherchée, à la fois par les scientifiques et au-delà par le grand public.

Deux grandes catégories d'usages utilisent la notion de workflow : les protocoles expérimentaux, dans des domaines tels que la biologie, l'astronomie, la physique, la neuroscience, la chimie, etc. (workflows scientifiques) et les chaînes de traitement

pratiquées dans des domaines commerciaux, financiers, pharmaceutiques (processus métiers). Elles donnent lieu à plusieurs pistes de recherche diverses, mais cependant connexes. Dans le cadre de cette thèse, nous traitons plus particulièrement les workflows scientifiques.

1.3.1.1 Définitions de workflow

La WfMC (Workflow Management Coalition) (WfMC, 1999) a donné une définition qui généralise la notion de workflow indépendamment des domaines spécifiques :

"Workflow is the automation of business process, in whole or part during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules."

Nous traduisons cette définition par : " Un workflow est l'automatisation d'un processus métier, en tout ou en partie, au cours de laquelle des documents, des informations ou des tâches sont passées d'un participant à un autre pour l'action, selon un ensemble de règles procédurales".

En ce qui concerne le workflow scientifique, nous retiendrons la définition suivante.

Un workflow est composé d'un ensemble de tâches (traitements) organisées selon un ordre logique, afin de réaliser un traitement global, complexe et pertinent sur un ensemble de données sources. Ces données sont souvent complexes, tant au niveau de leurs structures que de leurs organisations. Elles sont souvent volumineuses.

La taille d'un workflow scientifique peut varier de quelques tâches à des millions de tâches, qui sont souvent de calcul intensif "Computation Intensive" (Ludäscher, 2009). Pour les grands workflows, il est souhaitable de répartir les tâches entre plusieurs ressources, afin d'optimiser les temps d'exécution. En tant que tel, les workflows impliquent souvent des calculs répartis sur des clusters, des grilles, et d'autres infrastructures informatiques. Récemment, les clouds computing sont évalués comme une plateforme d'exécution de workflows (Hoffa, 2008 ; Juve, 2008). L'exécution d'un workflow est gérée par un SGWf (Système de Gestion de Workflow) dont l'architecture de référence est décrite brièvement dans la section suivante. 1

1.3.1.2 Les types de workflow

Regroupement des workflows : Une première manière de regrouper les workflows est la suivante :

1. les workflows documentaires (rédaction, validation, traduction de documents),
2. les workflows métiers, représentant un processus transverse à l'entreprise (concerne plusieurs entités organisationnelles). Il s'agit le plus souvent de procédures rattachées à une division opérationnelle de l'entreprise. Cette procédure peut être qualifiée de procédure métier, processus métier, procédure opérationnelle ou Business Process. Cela désigne un ensemble d'activités qui s'enchaînent de manière chronologique pour atteindre un objectif, généralement délivrer un produit ou un service, dans le contexte d'une organisation de travail (ex : une entreprise, administration, etc.).

Ainsi, on qualifie de Business Process (processus métier) l'ensemble des activités et de procédures qui permettent collectivement la réalisation d'un objectif métier. Le workflow correspond à l'automatisation de ce Business Process.

Critères	De production	Administratif	Ad-hoc	Collaboratif
Capacité de traitement	Haute capacité de traitement Temps de réponse rapide. Le but est la productivité	Capacité de traitement inférieure (10 à 100 fois moins que pour un workflow de production)	Facilite d'utilisation et d'apprentissage sont très importantes.	Capacité de changer dynamiquement la définition d'un processus est essentielle
Utilisation	Employés travaillant à plein temps sur des activités courtes.	Un grand nombre d'employés peuvent être impliqués	La modification dynamique et rapide des processus est essentielle.	Fournir une voie structurée pour travailler ensemble
Nature des processus	Processus formels avec peu de variation Les processus peuvent être très complexes.	Une variété de processus peut exister dans le même système. Les processus peuvent être bien définis, mais requièrent moins d'exigence.	Facilité de déploiement.	Les processus sont moins rigides

On peut aussi regrouper les workflows selon les catégories suivantes :

1. **le WorkFlow de production** : qui correspond à la gestion des processus de base de l'entreprise. Les procédures supportent peu de changements dans le temps, et les transactions sont répétitives (ex : contrats d'assurance, gestion de litiges, de réclamations clients, etc.),
2. **le WorkFlow Administratif** : gestion des procédures administratives par circulation de documents électroniques (ordres de mission, demandes de formations, ...) , automatise des processus variables (plusieurs cas) et mais bien définis (bien structurés).
3. **le WorkFlow Ad-Hoc** : généralement traitement de procédures secondaires (procédure d'exception) et dont la structuration est faible (circulation de notes d'information) et pour lesquelles il n'est pas toujours possible de définir des règles à l'avance, orientés vers le travail coopératif en groupe où l'initiative individuelle importante et problèmes à résoudre au cas par cas et en mode interactif.
4. **le WorkFlow collaboratifs** : Gestion d'un travail de groupe dont le processus est complexe, et souvent créé pour un travail de groupe particulier (travail collaboratif, équipe virtuelle,...) .

Comparaison entre types de workflows :

Chapitre 2

Modélisations et Réseau de Petri

Chapitre 3

Implémentation

Bibliographie

GOOSSENS, Michel, Frank MITTELBACH et Alexander SAMARIN (1993). *The \LaTeX Companion*. Reading, Massachusetts : Addison-Wesley.