Article publié dans la revue IT-EXPERT de septembre 2016

Métrologie applicative : Intrusif vs Non Intrusif, y a-t-il une voie à privilégier ?

Résumé

Le marché de la métrologie est en pleine effervescence et la richesse fonctionnelle des outils proposés permet de mieux maîtriser la qualité du service rendu par l'IT.

En outre, ces outils peuvent constituer le socle d'une offre de gestion de la qualité de service, commune à l'ensemble des acteurs de l'IT.

On privilégie en général à cet effet les approches non intrusives.

Toutefois, dans le domaine particulier des applications dites critiques (ou sensibles), nous estimons, selon nos retours d'expérience, que la voie d'une démarche « intrusive » doit être explorée en raison des bénéfices essentiels qu'elle apporte, sans toutefois prétendre qu'elle soit systématiquement l'unique réponse.

Les outils de métrologie du marché

APM, BAM, BTM, deep dive monitoring,… ? Autant d'acronymes et de solutions pour mettre en place une métrologie efficace dans un système d'information moderne.
L'objectif aujourd'hui n'est plus seulement de superviser les applications pour constater les dégradations et les dérives de temps de réponse ou de consommation de ressources.

Il devient essentiel de répondre à des questions de plus en plus complexes, comme par exemple :

- Quel est l'impact utilisateur d'une dégradation ? Quel est l'impact sur l'utilisabilité du service global ?
- Quelle(s) contremesure(s) mettre en place immédiatement ?
- Comment cerner l'origine de l'incident pour mobiliser les bonnes ressources ?

En pratique, et en partie parce que les architectures sont de plus en plus massivement distribuées (au sein même de l'entreprise ou entre partenaires), de nombreux incidents sont liés à des causes exogènes au service lui-même.

Outre l'incident technique, les causes peuvent être multiples, par exemple :

- Un succès beaucoup plus important que celui attendu pour une nouvelle fonction,
- Un événement externe qui modifie l'usage habituel du service (crack boursier, succès d'une campagne marketing, « buzz », attaque sécuritaire,…),
- Une mauvaise utilisation du service.

Les outils traditionnels de supervision des ressources de l'IT (CPU, mémoire, disque,…) ne sont plus suffisants pour répondre à ces questions. En effet, les ressources peuvent être toutes entièrement disponibles, sans que le service soit bien rendu pour autant. Inversement, les ressources peuvent être utilisées à 100% sans que cela n'ait d'impact significatif sur la qualité du service rendu.

La supervision active et échantillonnée

La première réponse, en complément à la supervision des ressources, a été de mettre en place une supervision « active » à base de robots simulant un utilisateur.
Mais cette réponse n'est plus suffisante puisque l'échantillonnage de ce flux de simulation est très réducteur en termes de représentativité quantitative (une transaction pour 10000, 100000 ?) mais également qualitative (la diversité des usages d'une application n'est, le plus souvent, pas simulable).

La supervision passive et exhaustive

Les solutions de dernière génération sont basées sur une supervision « passive », c'est-à-dire qu'elles analysent le flux réel des utilisateurs.

Elles mesurent la performance de tous les utilisateurs en s'appuyant sur une analyse des flux réseau, et/ou de l'exécution des applications grâce à des agents spécifiques.
Ceci permet d'avoir une vision exhaustive de la qualité du service rendu pour l'intégralité des utilisateurs réels.
Ces mesures sont déclinées selon plusieurs axes, ce qui permet de détecter des dérives sur des sous-éléments qui n'impactent pas la moyenne générale, et il est souvent possible de déclencher des alertes sur franchissement de seuil pour les transactions unitaires.

Gartner indique que les systèmes de capture de paquets et d'analyse, non intrusifs, sont utilisés par quelques-unes des solutions actuellement les plus efficaces du marché.

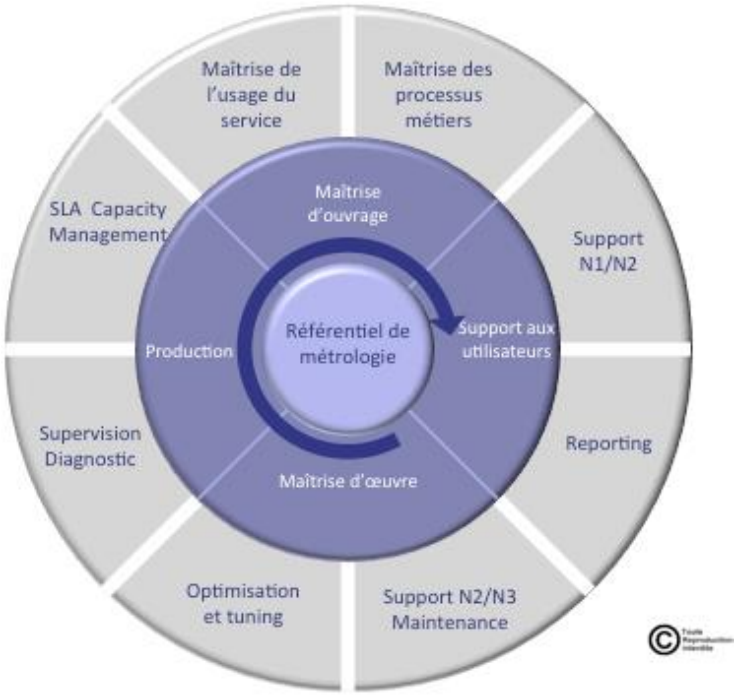
Les grandes familles de solutions

Le BAM (Business Activity Monitoring) regroupe la famille des outils qui fournissent en temps réel un résumé de la situation des activités métiers.

Derrière l'acronyme APM (Application Performance Management) se cachent les outils permettant de contrôler la performance des applications métiers et de détecter rapidement d'éventuels problèmes directement liés aux logiciels ou à l'infrastructure (réseau, système, base de données,...), alors que les logiciels de BTM (Business Transaction Management) s'attachent à suivre les transactions à tous les niveaux de l'infrastructure.

Les outils de deep dive monitoring permettent d'aller, comme le nom l'indique, plus profondément en proposant une vision de l'exécution du code d'une transaction, et sont plutôt à l'usage des développeurs dans le cadre de la recherche de diagnostics.

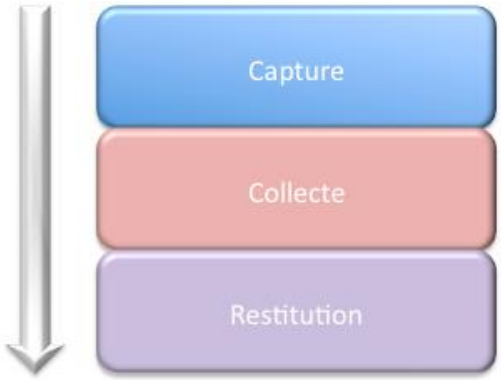
L'ensemble de ces outils adresse des besoins et des métiers différents, mais complémentaires, illustrés par le schéma suivant.



Comment fonctionne la métrologie applicative ?

Tous les outils du marché se basent sur les opérations suivantes :

- La capture des informations,
- La collecte, le stockage et l'indexation de ces informations,
- La restitution (corrélation, agrégation, alertes et mise en forme du reporting).



La capture

Les outils du marché fonctionnent en mode « non intrusif » vis-à-vis du code applicatif, c'est-à-dire que de façon générale, il n'est pas nécessaire d'intégrer des APIs de l'outil dans le code source des applications mesurées.
Ils sont cependant intrusifs vu de l'OS et/ou du réseau et/ou des middleware (JVM pour Java, CLR pour .Net, …).
La capture est alors très souvent réalisée à travers des composants logiciels spécifiques à chaque technologie (agent JVM, agent réseau, agent de parsing de logs…).
Ces agents permettent souvent de faire de la découverte (c'est-à-dire de décrire la topologie réseau), ils capturent les événements unitaires (appel de telle fonction, tel service) et mesurent le temps de réponse de chaque événement.

Afin de permettre d'associer l'ensemble des événements unitaires d'une transaction donnée, ces agents ajoutent un identifiant de corrélation à chacun d'entre eux, qu'ils véhiculent d'appel en appel.

L'enjeu essentiel pour chaque outil du marché est d'avoir l'empreinte la plus légère possible afin ne pas détériorer les performances de l'application ainsi mesurée.

Pour l'IT, les autres critères de choix sont :

- La couverture par ces agents en termes de technologies et middlewares supportés,
- La capacité à réaliser une mesure de bout en bout (depuis le navigateur de l'utilisateur, par exemple pour une application Web, ou dès l'entrée dans l'infrastructure IT).

La collecte et le stockage

Le rôle des collecteurs est de recueillir et de stocker de façon plus ou moins structurée les informations issues des agents dans un repository.

La restitution

L'accès à ce repository est réalisé par une interface de requêtage et de reporting souvent très riche et ergonomique.

On constate que :

- Ces outils sont totalement intégrés, de la capture à la restitution. Aussi, quand les différents besoins de restitution (APM, BPM,…) coexistent pour une même application, ce modèle oblige à plusieurs opérations de capture (une par outil).
- La plupart de ces outils ont un format de repository fermé, ce qui rend toute corrélation avec d'autres sources de données (comme par exemple les données de la supervision d'infrastructure, les données métiers, … et les données des outils de métrologie entre eux) difficile, voire impossible.

Même si la valeur ajoutée métier de ces outils est principalement celle créée par la restitution des informations, c'est le mode de capture propre à chaque outil qui limite la richesse fonctionnelle et l'extensibilité de cette restitution.

Puisque chaque outil n'adresse que son périmètre d'utilisation, quand les besoins de métrologie sont multiples (APM, BPM,…) et/ou plus larges (sécurité,…), on constate que le mode de capture « non intrusif » arrive à ses limites.

Le cas des applications critiques

Selon le CWA(1), une application critique (ou sensible) est une application dont la qualité de service est une exigence vitale. Toute dégradation du service a un impact direct sur le métier de l'entreprise.

Ces applications nécessitent le plus souvent conjointement :

- Une solution d'APM :
  - Mesure des temps réponse avec la finesse désirée (moyenne générale de l'application déclinée par fonction, sous-fonction,…),
  - Mesure d'impact d'une dégradation,
  - Gestion d'alerte sur franchissement de seuil (moyenne ou unitaire),
  - Aide au diagnostic des dégradations de performance,
  - Aide au capacity planning (vue essentiellement IT),
  - Usage du service (vue essentiellement IT),
  - Etc.
- Une solution de BPM :
  - Mesure des temps de traitement des business process,
  - Etat d'avancement des business process,
  - Aide au capacity planning (vue essentiellement métier),
  - Usage du service (vue essentiellement métier),
  - Etc.
- Une solution de deep dive monitoring :
  - Aide au diagnostic avancé sur incident ou dégradation de performance,
  - Etc.
- Une solution d'analyse sécuritaire :
  - Détection de déni de service,
  - Détection de comportements frauduleux,
  - Etc.
- Une solution de support aux utilisateurs finaux.

Par conséquent, pour couvrir l'ensemble de ces besoins, il y aura autant d'opérations et de formats de capture, de collecte et de restitution que d'outils mis en place.

Par exemple, pour ce type d'application, le support à l'utilisateur ne peut se contenter des seuils d'alertes fournis par les outils d'APM pour traiter tous les appels entrants.

En outre, l'interprétation des données issues des différentes captures peut induire des difficultés de compréhension entre les différents propriétaires de ces données :

- Le temps de réponse est une notion qui a une acception différente selon les acteurs. Aussi, un temps de réponse constaté au niveau de l'infrastructure peut être différent de celui ressenti par l'utilisateur (typiquement pour des raisons de latence au niveau du réseau).
- Les acteurs du marketing vont plutôt raisonner « appel de grande fonction et de page », alors que les acteurs de l'IT vont plutôt raisonner « requêtes et transactions unitaires ». Dans ces conditions, il est important de partager le même référentiel.

Et pourtant, les données unitaires nécessaires aux métiers, au support, à la sécurité, au « run », ou encore au diagnostic avancé sont-elles si fondamentalement différentes ? Ou est-ce plutôt la façon de les présenter et de les analyser qui est spécifique à chaque acteur ?

Bien entendu, certaines données sont spécifiques à un usage, et les règles de sécurité pour accéder à ces données peuvent également être différentes en fonction de celui-ci. Mais il n'y a rien de bien révolutionnaire pour une application que d'accéder à ces données en fonction de profils et de droits.

Un service de banque à distance a, par exemple, les besoins suivants :

- Assistance aux diagnostics : chaque incident (et pas uniquement ceux liés à la performance) doit être résolu au plus vite et même anticipé.
- Assistance au support aux utilisateurs : par exemple pour pouvoir répondre à une question telle que « je ne comprends pas, j'ai vu sur mon écran des données qui ne m'appartiennent pas ! ».
- Assistance au service en charge de la sécurité : analyse comportementale pour détecter les utilisateurs victimes de phishing, de keylogger,…
- Assistance au marketing : comment mes clients naviguent-ils dans l'application, est ce qu'une fonction a du succès, quels navigateurs sont utilisés, …
- …
- Sans oublier des besoins très ponctuels liés à l'actualité.

Sur la base d'un tel constat, le responsable de l'IT est en droit de se poser les questions suivantes : Faut-il multiplier les outils pour couvrir tous ces besoins ? Une autre voie est-elle possible ?

Une autre voie est-elle possible ?

Les limites induites par le mode de capture en mode « non intrusif » peuvent être levées par une approche intrusive.
Par intrusif, on entend l'intégration, dans le code source de l'application, des APIs nécessaires à la capture des informations désirées.
Cela demande donc de disposer du code source, ce qui en pratique exclte les progiciels.
Cependant, certains middleware et progiciels disposent nativement de systèmes de « logs » ou même de monitoring qui permettent de capturer les informations nécessaires.

La capture n'est pas une fin en soi et c'est au niveau de l'analyse et de la restitution que se situe la valeur ajoutée. Or, l'émergence de solutions industrielles et ouvertes (permettant de construire des solutions couvrant les besoins de collecte et de restitution) rend aujourd'hui le mode de capture intrusif possible et puissant.

Les avancées de ces dernières années les plus remarquables concernent la collecte, le stockage et la restitution intelligente de ces données, notamment :

- Les Event Streaming Processing (ESP) ou Complex Event Processing (CEP), pour la corrélation, l'agrégation, l'application de règles, …
- L'indexation « plain text », pour la recherche, la corrélation, …
- Le reporting, pour l'élaboration de consoles et de rapports hautement configurables.
- L'analyse multi dimensionnelle (SQL ou autre).
- Etc.

Il existe de nombreux outils qui adressent tout ou partie de ces besoins et il est possible de les faire inter-opérer afin d'extraire la « substantifique moelle » de ces informations.

Bien évidemment, la mise en place d'un tel mode de capture intrusive nécessite de définir les informations unitaires dont on a besoin et la façon de les capturer. Elles sont schématiquement de trois natures :

- Techniques : toute fonction à l'origine de latence (appel d'un webservice, d'une base de données, …) doit être chronométrée. Le taux d'occupation des pools doit être également calculé, …
- Fonctionnelles : les fonctions métiers, leur statut (succès, échec avec la cause), et les attributs métiers associés sur lesquels on veut faire des recherches, des agrégations, …
- Sécuritaires : accès aux fonctions sensibles avec application des règles de sécurité (confidentialité, intégrité, preuve).

De plus, il est nécessaire de mettre en place un système de corrélation qui permette de réassocier entre elles, et dans tout le système, ces informations capturées.

Ces informations constituent le minimum de traçabilité qu'on est en droit attendre d'une application critique afin d'en maîtriser la qualité du service rendu (voir les normes de qualité logicielle décrites dans le CWA[1].

La charge de travail pour mettre en place cette capture dépend des applications. :

- Si cette capture est anticipée dès la phase de conception, cette charge sera faible.
- Dans les autres cas, cela dépend en particulier du niveau d'encapsulation de l'application. Concrètement, on trouve souvent des points de passage obligés et centraux dans le code applicatif pour capturer les informations. En revanche, la transmission des identifiants de corrélation peut être plus problématique lorsqu'elle n'est pas anticipée.

Le mode de capture intrusif associé aux outils de collecte et de restitution (apparus ces dernières années) apporte l'agilité indispensable à la prise en compte rapide et efficace de tous les besoins liés aux applications critiques, y compris les besoins spécifiques tels que l'analyse comportementale de sécurité.

Certains ont-ils mis en œuvre cette voie intrusive ?

La réponse est oui, et cela peut même aller jusqu'à asservir le comportement de l'application afin d'éviter des conditions de déni de service, par exemple en suspendant une

fonction qui, pour une cause externe (flux vers un partenaire), subit un ralentissement pouvant avoir un impact plus large.

En conclusion

Le choix d'une voie non intrusive ou intrusive reste une question ouverte sachant que l'une n'exclut pas l'autre au sein d'un même système d'information.
Dans tous les cas, la voie intrusive doit au moins être explorée pour les applications critiques compte tenu des bénéfices qu'elle peut apporter.

Rédacteur :
Sébastien BARNOUD
Directeur technique et co-fondateur de PROLOGISM
www.prologism.fr

(1) CWA = CEN Workshop Agreement "Best Practices for the Design and Development of Critical Information Systems"

Référentiel issu d'un groupe de travail européen :

- qui s'est constitué à l'initiative de PROLOGISM sous l'égide du CEN* et avec le concours méthodologique de l'AFNOR,
- auquel se sont joints des acteurs du marché des SI critiques (CS, La Banque Postale, NYSE EURONEXT, THALES, Groupement Cartes Bancaires ...),
- dont PROLOGISM a assuré la présidence durant les deux années qui furent nécessaires à l'élaboration du référentiel.

Ce document est désormais du domaine public sous le label CWA, conféré et promu par le CEN

(*) CEN : European Committee for Standardization – Comité Européen de Normalisation – Europäisches Komitee für Normung