Incident Terraform (incident #3)

Date : du 16/01/2017 17:34 au 16/01/2017 20:34

Auteurs du PostMortem

- Quentin CATTEZ (CQU)
- Adrien SAUNIER (ASA)
- Thomas WICKHAM (TWI)

Statut : RÉSOLU

Résumé

En utilisant le playbook Ansible, la tâche déployant la configuration (“Terraform apply”) a commencé à être exécuté suite à la tâche définissant la nouvelle configuration (“Terraform plan”). Il fallait interrompre le playbook au bon moment pour relire le plan, et le moment a été trop tardif.
L’action du playbook a détruit toutes les machines (et les aurait recréées si on n’avait pas interrompu le playbook).
Toutes les machines de PROD ont été détruites dont la BDD. Le site est DOWN. Pas de mode maintenance possible car ce mode est porté par les frontaux Web Nginx (détruits).

Impact

L’application est DOWN. Toutes les machines ont été détruites. Perte de 1j de données. Perte du dashboard de monitoring.

Causes Originelles

- L’ergonomie du playbook Ansible qui modifie le provisionning de la PROD, potentiellement destructeur (on doit faire Ctrl+C pour lire le plan)
- Terraform qui détruit toutes les machines même si la seule modification effectuée consistait à ajouter une machine.

Évènement déclencheur: provisionning d'une machine avec Ansible

Résolution: 16/01/2018 20:34

- Reconstruction des machines
- Redéploiement total des configurations
- Restauration manuelle du dernier backup de la BDD
- Redéploiement des applis

Détection

- Thomas a lu l'output du playbook et constaté dans le plan Terraform que les machines n'avaient plus les mêmes IPs
- Notification immédiate sur Slack que Pix est down par UptimeRobot

Points d'Action

Point d'action	Type (résolution, atténuation, prévention)	Propriétaire de l'action	Tracking
Réparer la réplication et le backup de la DB sur PG Slave	résolution	TWI + CQU	https://trello.com/c/uXzsE1Bl/1032-r%C3%A9parer-la-r%C3%A9plication-de-la-base-de-donn%C3%A9es-et-la-proc%C3%A9dure-de-sauvegarde
Ré-importer les dashboards Grafana	résolution	TWI + CQU	https://trello.com/c/bdER6dob/327-importer-les-dashboard-grafana
Modifier le playbook Ansible pour forcer une validation manuelle des modifications d'infra	prévention		https://trello.com/c/jenoNso5/318-modifier-le-playbook-ansible-terraform-pour-%C3%A9viter-de-d%C3%A9truire-la-prod
Augmenter la fréquence des backups DB (toutes les heures)	atténuation		https://trello.com/c/Uxz854IQ/320-augmenter-la-fr%C3%A9quence-des-backup-de-la-bdd-de-production

Ajouter un backup automatisé de Grafana	atténuation		https://trello.com/c/eg8yhUJV/194-automatiser-les-backups-grafana
Avoir un Environnement de SANDBOX pour tester les modifications Terraform	prévention		https://trello.com/c/LxZvo3Fa/323-cr%C3%A9er-un-environnement-de-sandbox-pour-tester-les-modifications-terraform
Voir si les snapshots OVH peuvent nous aider	atténuation		https://trello.com/c/gjZMHkl9/321-voir-ce-que-les-snapshots-ovh-permettent-en-cas-de-perte-des-machines
Avoir un playbook Ansible qui dump les IPs des machines	atténuation		https://trello.com/c/IN5worT6/322-cr%C3%A9er-un-playbook-ansible-qui-dump-les-ips-des-machines
Avoir un playbook ansible qui configure automatiquement les DNS	atténuation		https://trello.com/c/Wjg9l87Z/324-cr%C3%A9er-un-playbook-ansible-qui-configure-les-dns-automatiquement
Avoir un script qui restaure les données	atténuation		https://trello.com/c/KeG2hNdr/325-cr%C3%A9er-un-script-et-une-t%C3%A2che-qui-restaure-la-bdd-si-vide

Leçons retenues

Ce qui s'est bien passé

- La résolution de l'incident sans trop d'encombres en 3h
- Le backup et la restauration des données
- Notifications Slack / UptimeRobot
- Le playbook Ansible (deploy-infra.yml)
- La communication avec Benjamin, Benoît, Nathalie

Ce qui a raté

- Le middleware de réplication PostgreSQL (RepMgr) qui est passé de la version 3 à 4 avec des breaking changes à corriger en live
- Grafana non sauvegardé

Là où on a eu de la chance

- Pas de certifications en cours
- La connaissance de TWI sur Postgresql : les rôles, databases
- La récupération des infos sur OVH par ASA qui a accéléré la résolution

Chronologie

16/01/2018 (toutes les heures (Heure de Paris))

- 17:34 : L'action Terraform "apply" du playbook Ansible a été jouée
- 17:34 : **Destruction** des machines de PROD
- 17:37 : Création du channel Slack incident-16-01
- 17:41 : Création du Post-Mortem
- 17:45 : Execution des actions "plan" et "apply" Terraform
- 17:54 : Déploiement des Nginx : `ansible-playbook -i inventory/pix_production deploy-infra.yml --limit nginx`
- 17:57 : Récupération des IPs des Nginx sur OVH Horizon pour reconfigurer le LoadBalancer OVH
- 17:57 : Passage en mode maintenance : `ansible-playbook -i inventories/pix-production activate_maintenance_mode.yml`
- 17:57 : Un simple curl en localhost ne fonctionne pas
- 18:00 : Connexion avec le compte de Jérémy sur OVH
- 18:05 : Ajout des IPs des Nginx au LoadBalancer (Cloud/Load Balancer), application de la configuration
- 18:06 : **Mode Maintenance ON grâce aux précédentes actions**
- 18:09 : Exploration de la piste d'ajout de volumes pour les BDDs
- 18:24 : Accord trouvé sur la non utilisation des volumes pour l'instant (impact trop important)
- 18:32 : Modification manuelle de la configuration des DNS avec les nouvelles IPs des machines
- 18:38 : Déploiement de la nouvelle configuration des DNS
- 18:40 : Début de modification de la configuration de Repmgr (options obsolètes)
- 19:06 : Déploiement de la configuration pour le reste des machines
- 19:13 : Redéploiement sur les hosts PG (problème avec la conf repmgr)
- 19:31 : Configuration du PG master OK. Configuration du PG slave NOK
- 19:44 : Installation de swift sur la machine PG master pour télécharger le backup
- 19:55 : Création manuelle de la base de données pix_production

- 20:13 : Restauration OK et nettoyage historique après opérations manuelles sur PG master
- 20:34 : Désactivation du mode Maintenance => **Résolution de l’incident**

Supporting information