re:Invent 2019 : l'IA à la mode AWS

Promesse tenue pour AWS.

La branche cloud d'Amazon s'était donné jusqu'à fin 2019 pour annoncer la disponibilité générale de son offre Outposts.

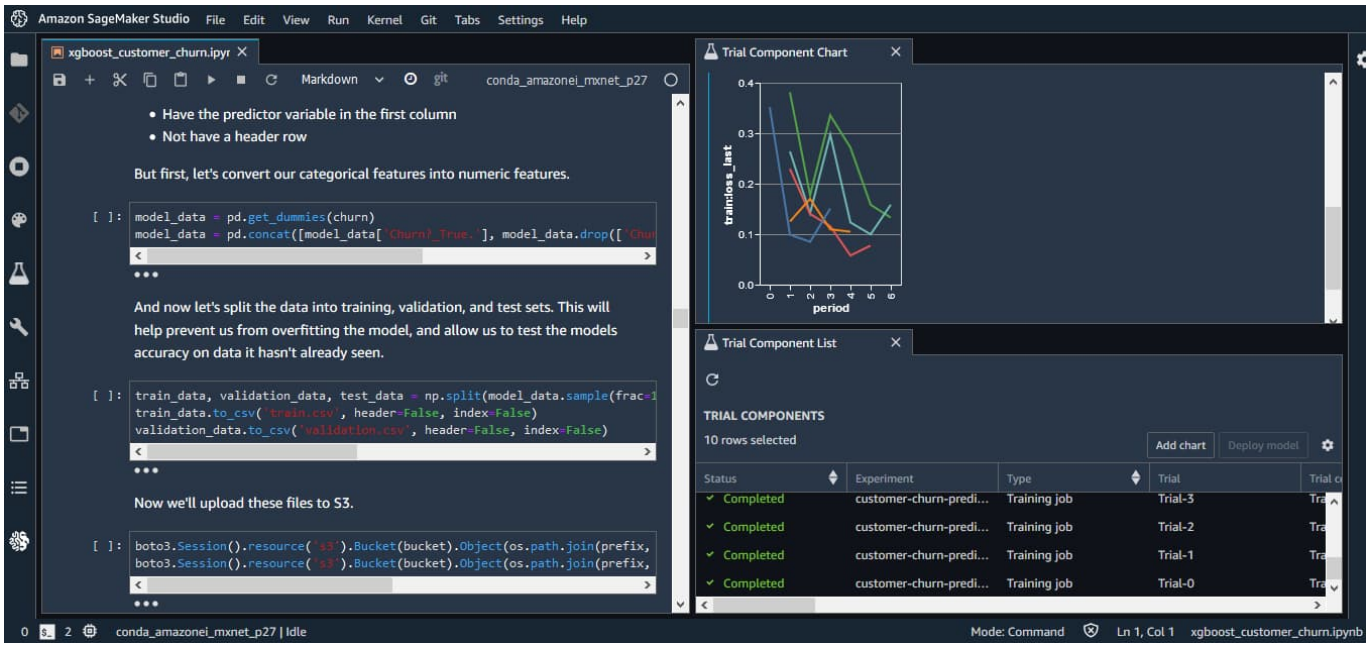
C'est chose faite, avec toujours la même promesse : permettre – comme Microsoft le fait avec Azure Stack – d'exploiter les capacités du cloud public sur des infrastructures internes. Dans le viseur, les charges de travail qui nécessitent une faible latence ou des traitements de données locaux.

AWS a profité de sa conférence re:Invent pour officialiser d'autres lancements. Par exemple, l'intégration d'EKS et de Fargate. C'est-à-dire de son service Kubernetes managé et de son moteur d'exécution de conteneurs sans gestion de serveurs ni de clusters.

Un synthé pour l'IA

De nombreuses annonces ont eu trait, de plus ou moins près, à l'IA. Avec en tête de liste l'outil SageMaker, destiné à la conception d'algorithmes d'apprentissage automatique. Lui sont ajoutés, entre autres :

- Un opérateur Kubernetes pour créer des tâches d'entraînement à partir des outils natifs de l'orchestrateur
- Autopilot, pour automatiser la création de modèles de classification et de régression
- Model Monitor, pour faciliter la supervision des modèles en production et détecter les problèmes de qualité de données
- La bibliothèque *open source* Deep Graph (basée sur PyTorch et MXNet) pour implémenter des GNN (Graph Neural Networks, capables d'analyser des données structurées en graphes)
- Un débogueur et un environnement de développement (illustré ci-dessous)



AWS étend aussi sa gamme de produits électroniques destinés à familiariser les développeurs avec les différentes branches de l'IA.

En 2017, il y avait eu la caméra DeepLens, orientée sur la formation à l'apprentissage profond sous l'angle de la vision par ordinateur. La v2, lancée cette année, est disponible à l'achat en France.

Le catalogue s'était agrandi l'an dernier avec DeepRacer, une voiture de course au 1/18 pour les projets d'apprentissage par renforcement. La voilà mise à jour avec un lidar et une caméra stéréoscopique.

Le petit nouveau de cette année s'appelle DeepComposer. Il s'agit d'un synthé axé GAN (réseaux antagonistes génératifs). Prix annoncé : 99 $.

Puces dédiées

Toujours sur le volet IA, on notera le lancement d'instances EC2 sur base Cascade Lake qui exploitent les puces Inferentia d'AWS. Elles sont dédiées aux tâches d'inférence (établissement de prévisions à partir d'un modèle entraîné). Performances annoncées : jusqu'à 128 teraOPS sur des entiers 8 bits.

Instance Name	Inferentia Chips	vCPUs	RAM	EBS Bandwidth	Network Bandwidth
inf1.xlarge	1	4	8 GiB	Up to 3.5 Gbps	Up to 25 Gbps
inf1.2xlarge	1	8	16 GiB	Up to 3.5 Gbps	Up to 25 Gbps
inf1.6xlarge	4	24	48 GiB	3.5 Gbps	25 Gbps
inf1.24xlarge	16	96	192 GiB	14 Gbps	100 Gbps

Également sur EC2, on aura relevé le lancement de l'outil Image Builder, pour automatiser la gestion des images système.

Pour ce qui est des puces « made in Amazon », la deuxième génération des Graviton est officielle. Gravées en 7 nm, elle embarquent des cœurs Arm Neoverse 64 bits. AWS les intègre dans trois types d'instances qui proposent de 128 à 512 Go de mémoire (celle à 256 Go est accessible en bêta).

Bêta également, mais uniquement aux États-Unis, pour UltraWarm.

Il s’agit d’un niveau supplémentaire de stockage adossé à Elasticsearch Service. Il se place en deçà du niveau « hot », à un prix moins élevé, pour des données « semi-chaudes ».

AWS se positionne par ailleurs sur la fin de vie de Windows Server 2008 (14 janvier 2020). Ce avec l’offre EMP (End-of-Support Migration Program). Avec l’appui de ses partenaires, le groupe américain propose d’aider à migrer vers son cloud des applications exécutées sur d’anciennes versions de Windows Server que Microsoft a abandonnées.