ECOSYSTEMS DU BIG DATA DANS LES PLATFORMES CLOUD

11/22/2020

GCP, AWS, MICROSOFT AZURE, IBM CLOUD

Réalisé par :

Riali Mouad

Addi Kamal

Encadré par :

Pr . D.Zaidouni

Table de matières :

I- Introduction :

Le cloud computing permet de stocker, de traiter et d'analyser des données de manière plus évolutive, flexible et rentable qu'avec un déploiement sur site. Il offre également plus de sécurité. Lorsque les volumes de données connaissent une croissance exponentielle, ces avantages font toute la différence.

Le cloud offre plusieurs ressources de stockage et de traitement adaptées à des besoins d'entreprises, et cette dernière peuvent ainsi exploiter toute la valeur de ses données. En outre, pour les entreprises qui se lancent dans l'analyse big data, gérer des systèmes de big data sur site peut s'avérer extrêmement complexe. Avec le cloud, ces entreprises pouvent expérimenter ces besoins via des platformes tels que Google Cloud Platforme, AWS Amazon, Microsoft Asure et IBM Cloud.

II- Architectures du big data avec Google Cloud Platform (GCP) :

Google propose sur sa plateforme Google Cloud des services de plus haut niveau, **dédiés aux architectures du big data**. Les services Big Data de Google permettent notamment de traiter et d'analyser des données. Google BigQuery permet par exemple d'effectuer des requêtes sur des ensembles de données de plusieurs terabytes. Cloud Dataflow est une service de traitement de données conçu pour l'analyse, l'extraction, la transformation et le chargement de données. Cloud Dataproc offre des services **Apache Spark** et **Hadoop** pour le traitement Big Data. Elle intègre également les bases de données de Cassandra ou encore MongoDB.

Le figure suivant montre les différentes services dédiée au big data, offertes par GCP :

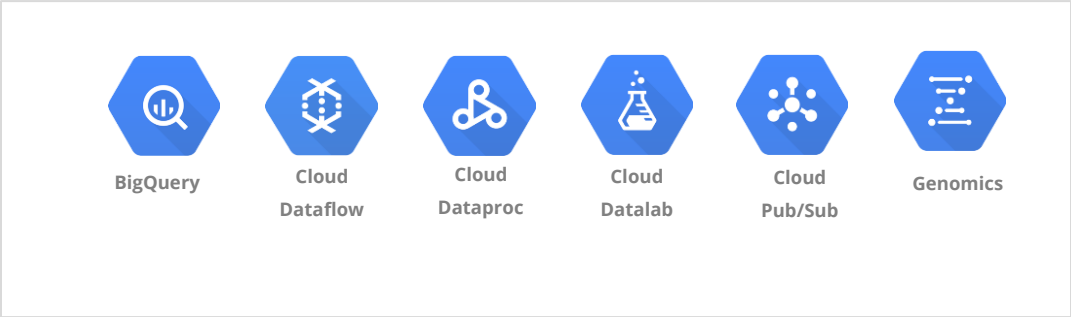


Fig1 : *Google Cloud Platform products and services for big data.*

Google Cloud Platform met tous ces produits Big Data à la disposition des entreprises. Il est possible de les utiliser depuis Google Compute Engine, Google Container Engine et App Engine. l'utilisateur peut télécharger n'importe quelle image Linux et exécuter son code dans la machine virtuelle sur Google Cloud Platform. Il peut aussi utiliser des conteneurs comme Docker par exemple.

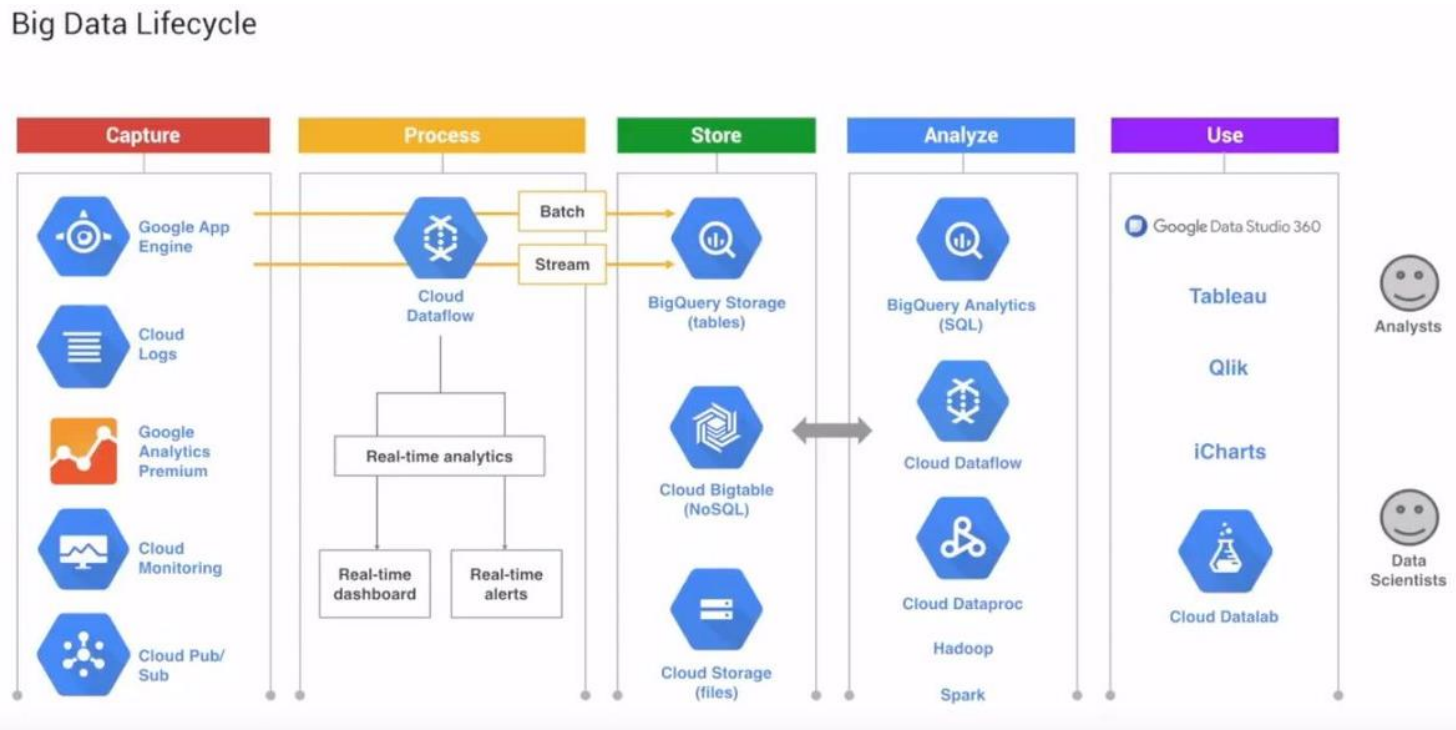
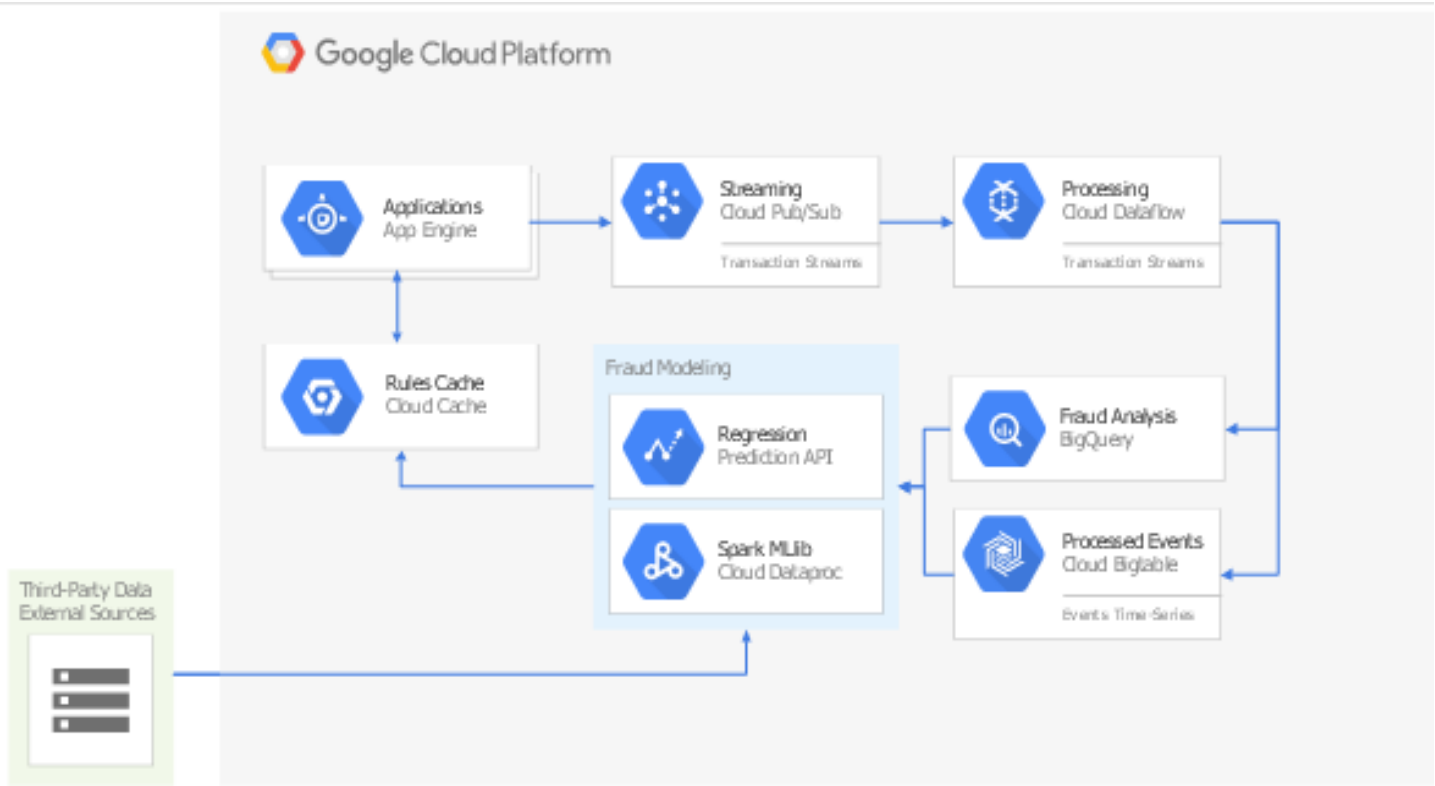
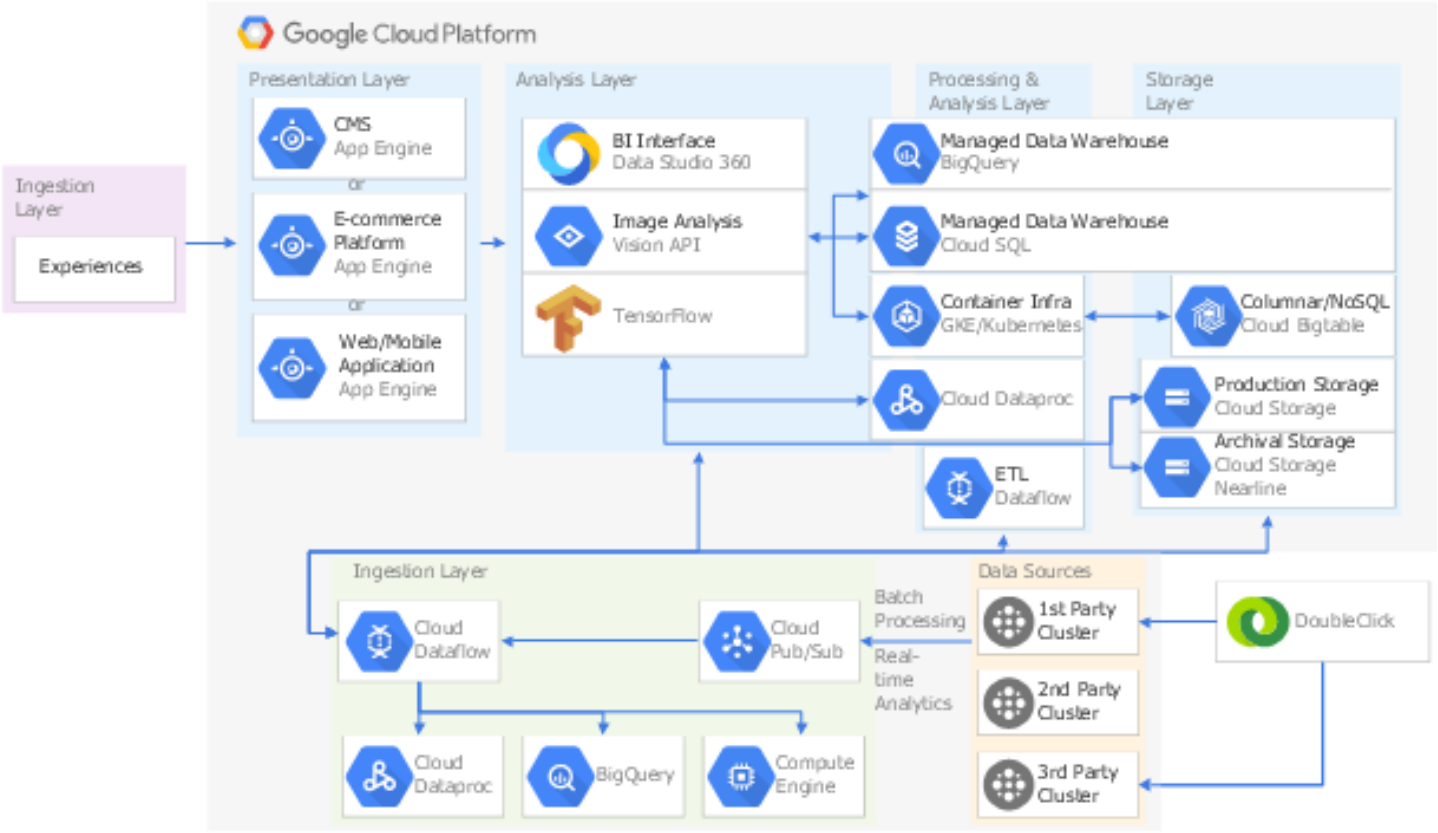


Fig2 : le cycle de vie d'une architecture big data.

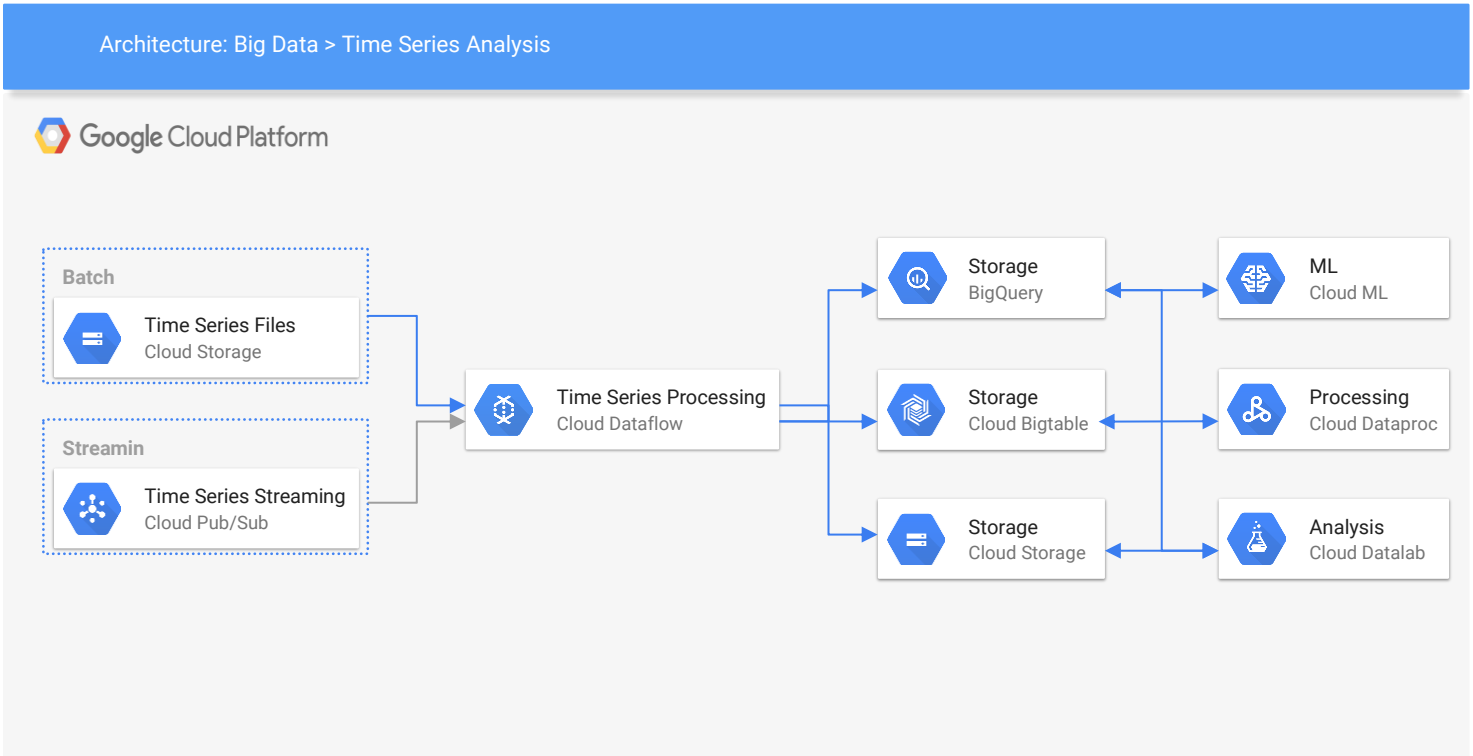
1- Detection ds fraudes :



2- Digital marketing :



3- Services financiers et analyse de séries temporelles :

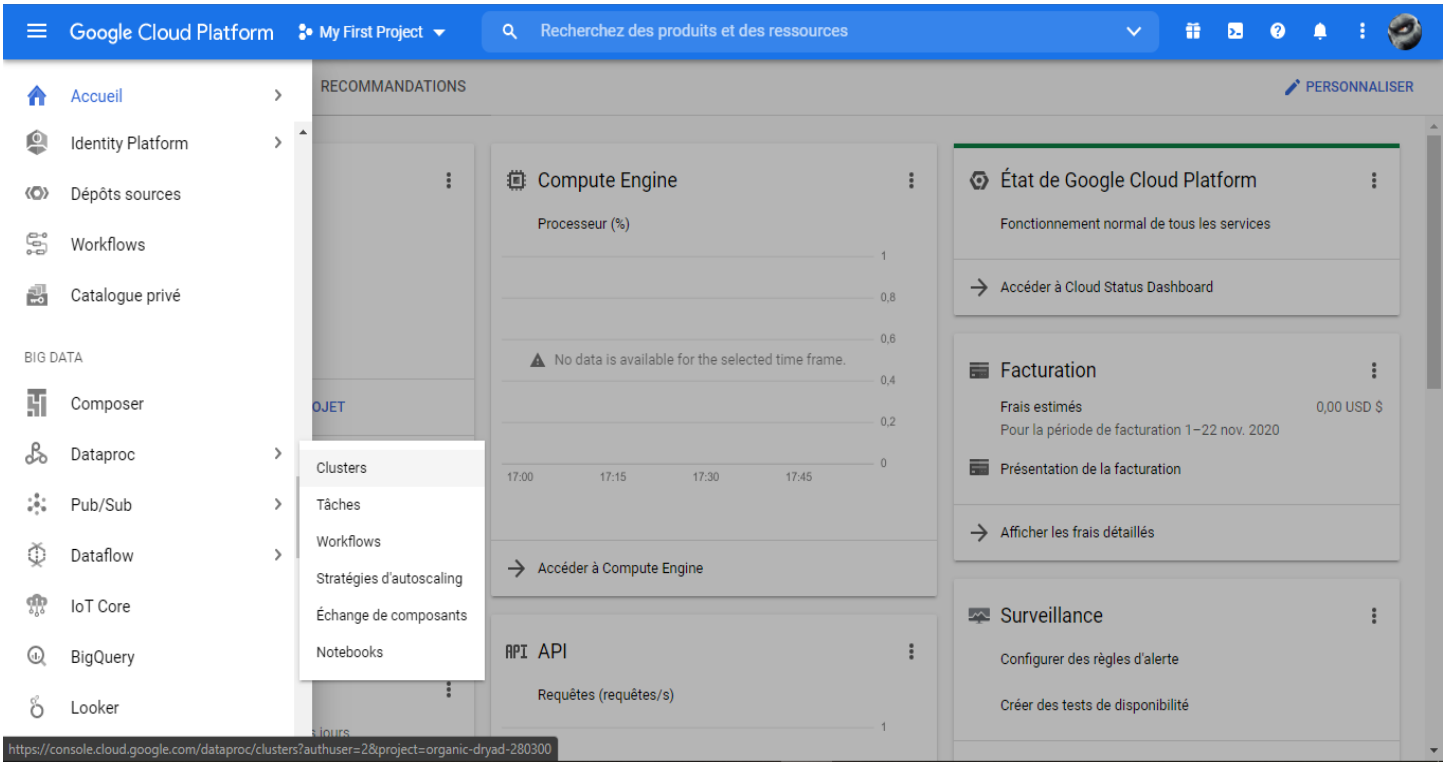


Dataproc :

Cloud Dataproc est un service cloud rapide, facile à utiliser et entièrement géré. En utilisant Dataproc dans GCP, nous pouvons exécuter des clusters Apache Spark et Apache Hadoop sur Google Cloud Platform de manière puissante et rentable.

Cloud Dataproc est le meilleur pour les environnements qui dépendent de composants spécifiques de l'écosystème Big Data.

Pour créer un cluster multi-nodes sur GCP avec Dataproc : selectionné l'onglet dataproc puis clusters :



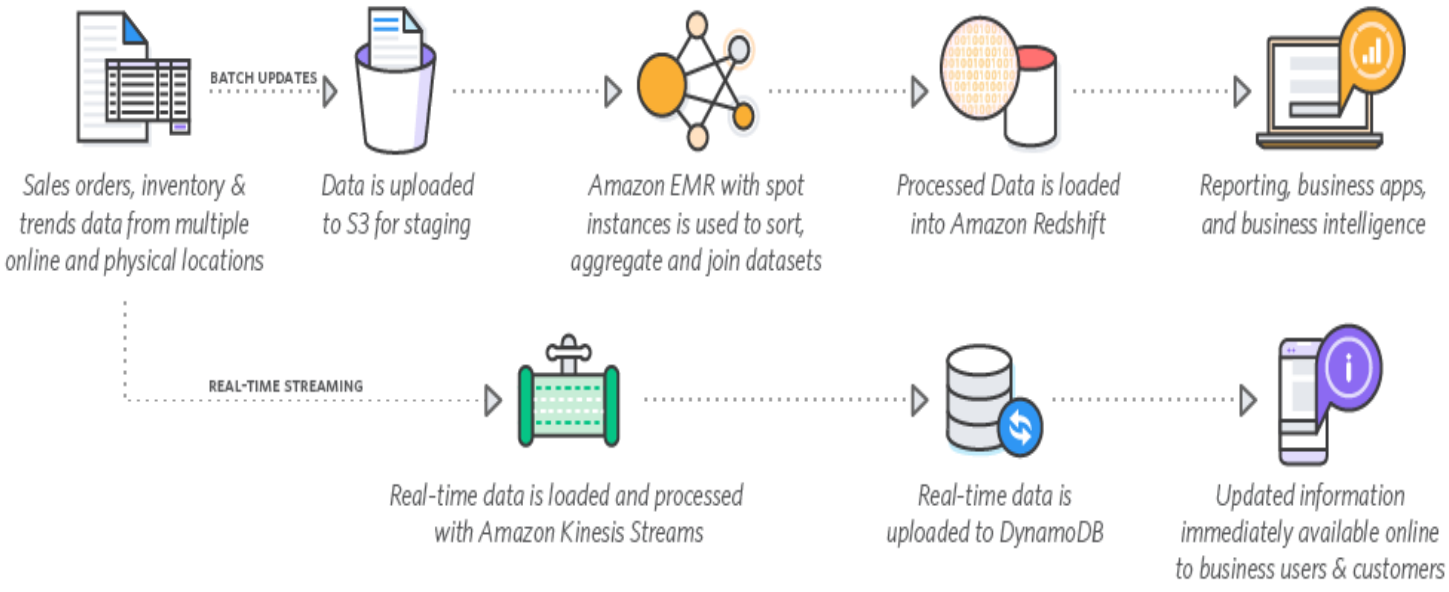
Une fenêtre de configuration s'ouvre sur lequel on choisit le nombre de neouds, memoires, CPU etc..

III- Architectures du big data avec AWS :

Amazon EMR est une plateforme leader de Big Data dans le cloud AWS dédiée au traitement de grandes quantités de données à l'aide d'outils à code source libre tels que Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi et Presto. EMR vous permet d'exécuter des analyses à l'échelle des pétaoctets à des coûts inférieurs de moitié à ceux des solutions sur site traditionnelles et à une vitesse trois fois plus rapide que celle d'un outil Apache Spark standard. Pour des tâches de courte durée, vous pouvez lancer et arrêter des clusters et payer suivant une tarification à la seconde pour les instances utilisées. Pour les charges de travail de longue durée, vous pouvez créer des clusters hautement disponibles que vous pouvez mettre automatiquement à l'échelle pour répondre à la demande. Si vous avez des déploiements sur site existants d'outils à code source libre, par exemple Apache Spark et Apache Hive, vous pouvez également exécuter des clusters EMR sur AWS Outposts.

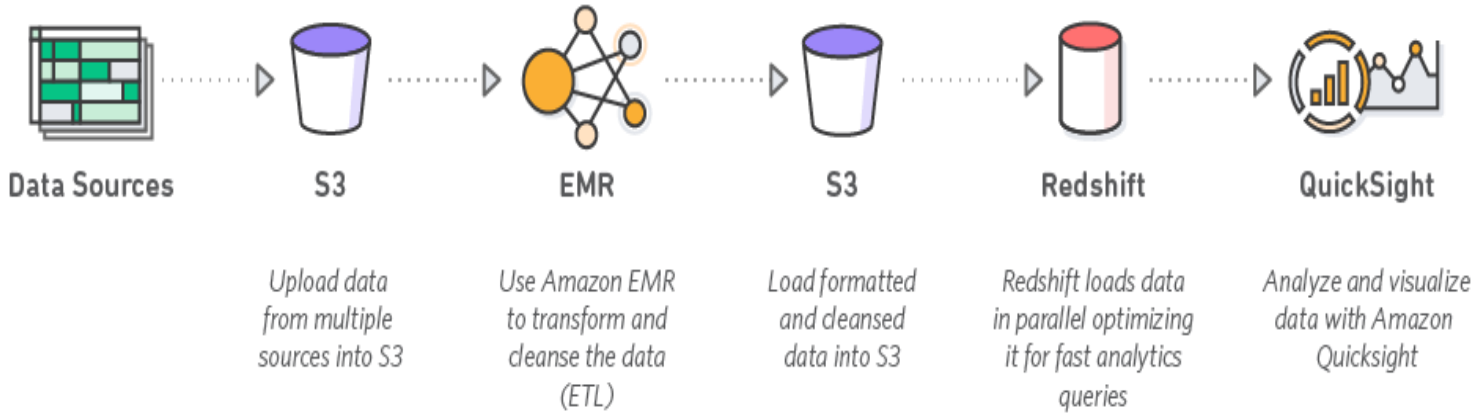
1- Analyses big data à la demande :

Avec AWS vous pouvez créer une application d'analyse complète qui dynamisera votre entreprise. Faites évoluer un cluster Hadoop de zéro à des milliers de serveurs en seulement quelques minutes, puis désactivez-le lorsque vous avez terminé. En d'autres termes, vous pouvez traiter les charges de travail Big Data plus rapidement et à moindre coût.



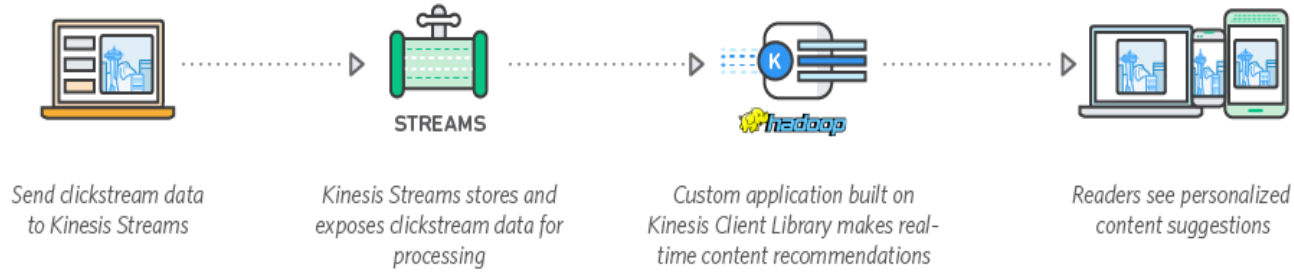
2- Entreposage de données :

Optimisez la performance des requêtes et faites des économies en déployant votre architecture d'entreposage des données dans le cloud d'AWS.



3- Analyse des parcours de navigation :

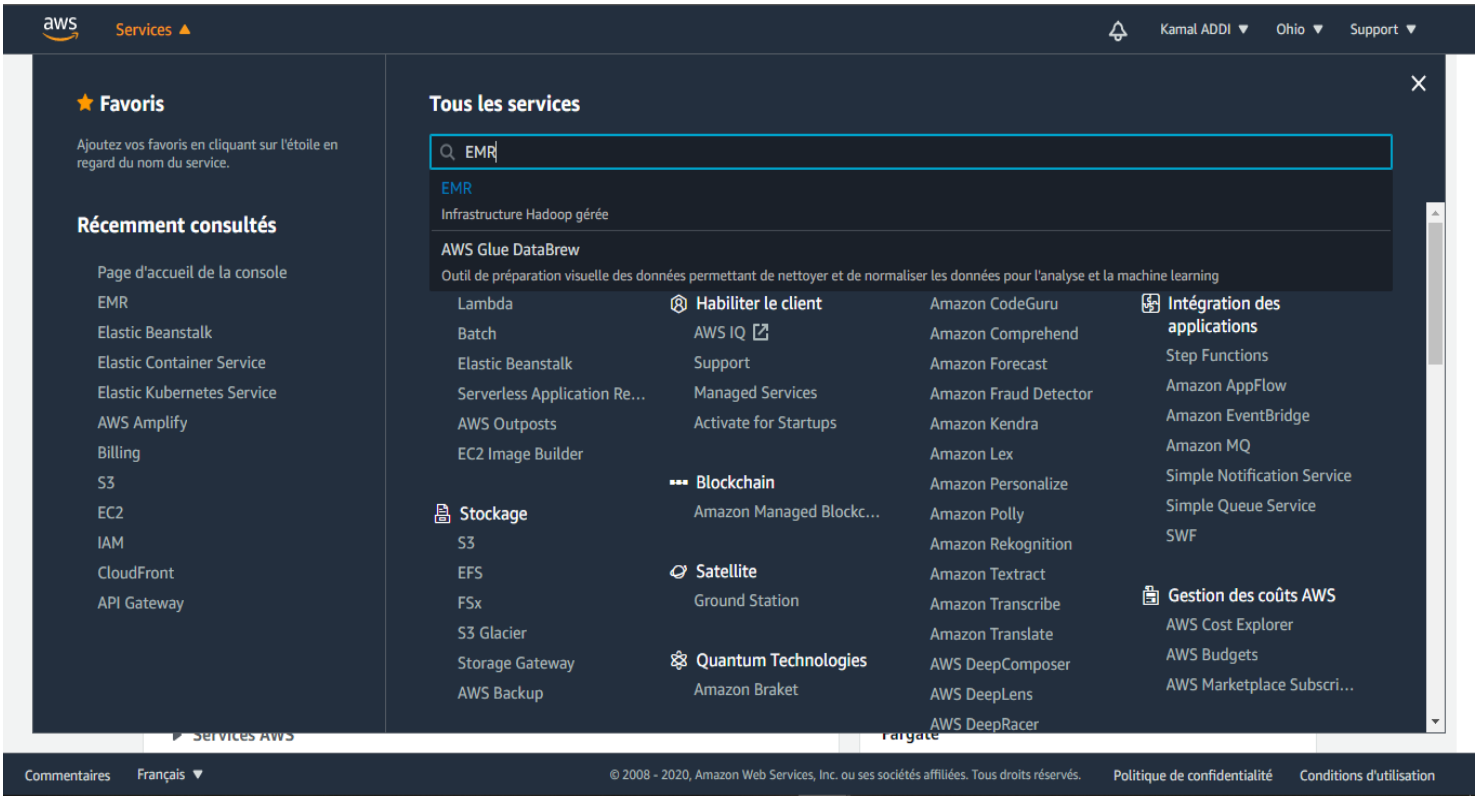
Améliorez l'expérience numérique de vos clients et apprenez à mieux connaître votre site Web. Collectez, traitez, analysez et visualisez des analyses de parcours de navigation en temps réel avec AWS.



- **Lancer un cluster Amazon EMR avec AWS :**

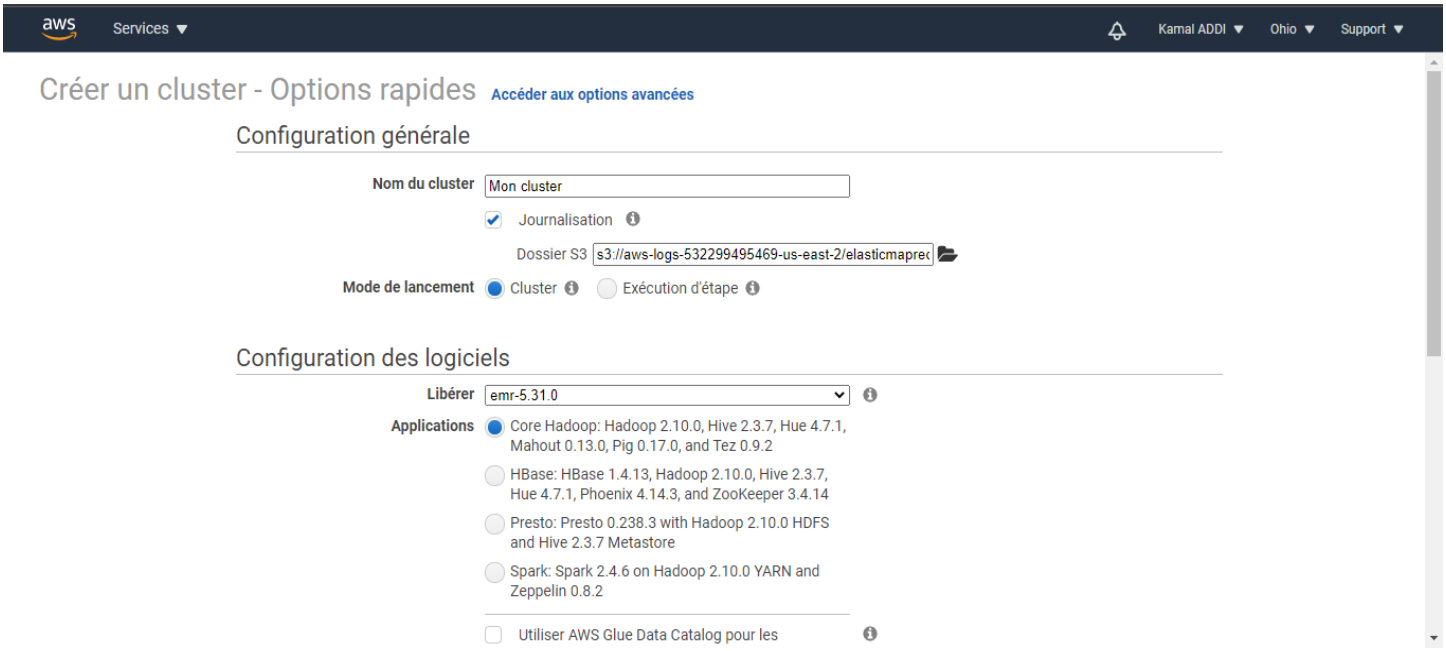
On peut lancez un exemple de cluster à l'aide des **Quick Options (Options rapides)** dans la console Amazon EMR, en qulques cliques :

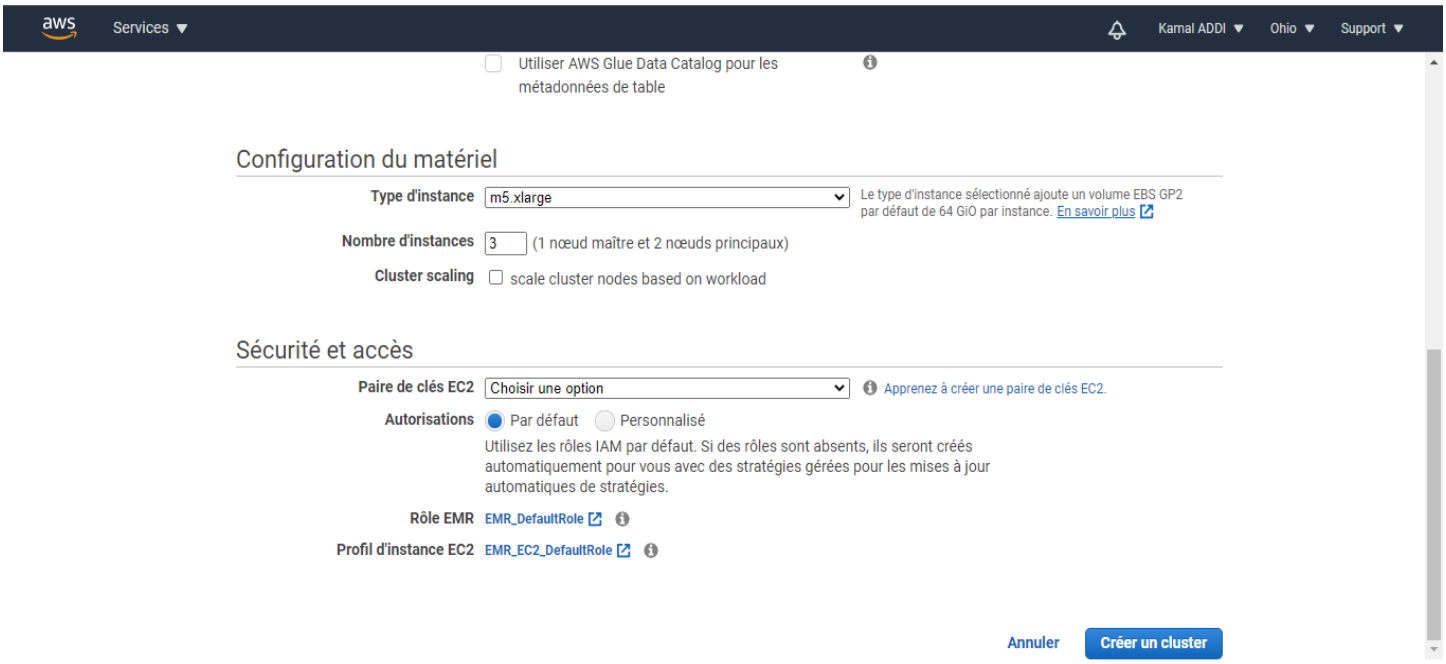
Dans la console AWS Management Console dans l'onglet services >> barre de recherche on tape EMR :



La page suivant s'ouvre : Choisissez **Créer un cluster**.

On remplit les champs suivants puis **Créer un cluster :**





La page d'état du cluster **Récapitulatif** s'affiche. On peut utiliser cette page pour surveiller la progression de la création du cluster et afficher les détails sur le statut du cluster. Les éléments sur la page de statut sont mis à jour une fois les tâches de création de cluster terminées.

IV- Architectures du big data avec Microsoft Azure :

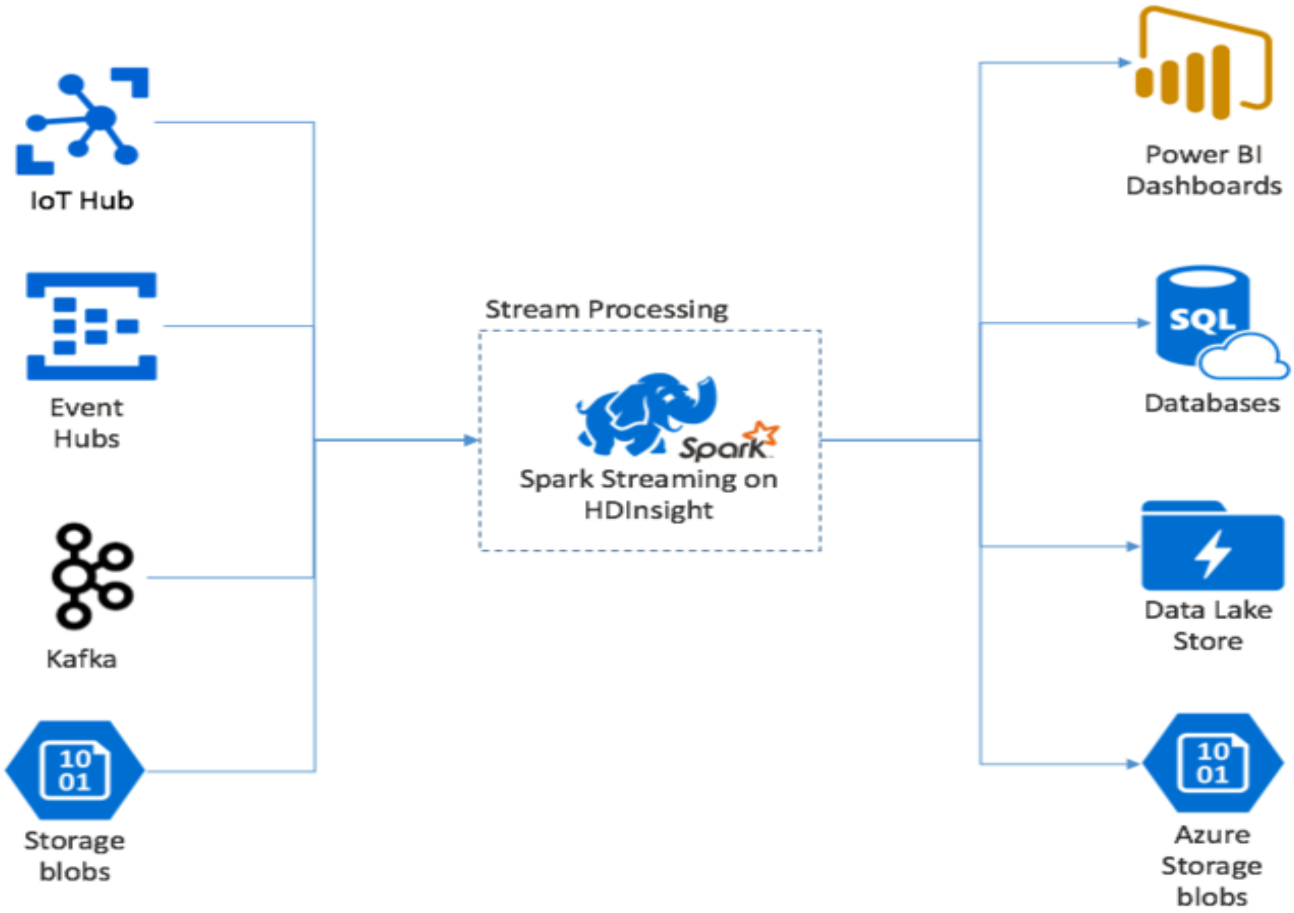
Comme la plupart des platformes cloud, Microsoft Azure a aussi des outils disponibles sur le marketplace pour la réplication et l'ingestion des données en temps réel.

Les solutions big data proposées par Microsoft sont les suivantes :

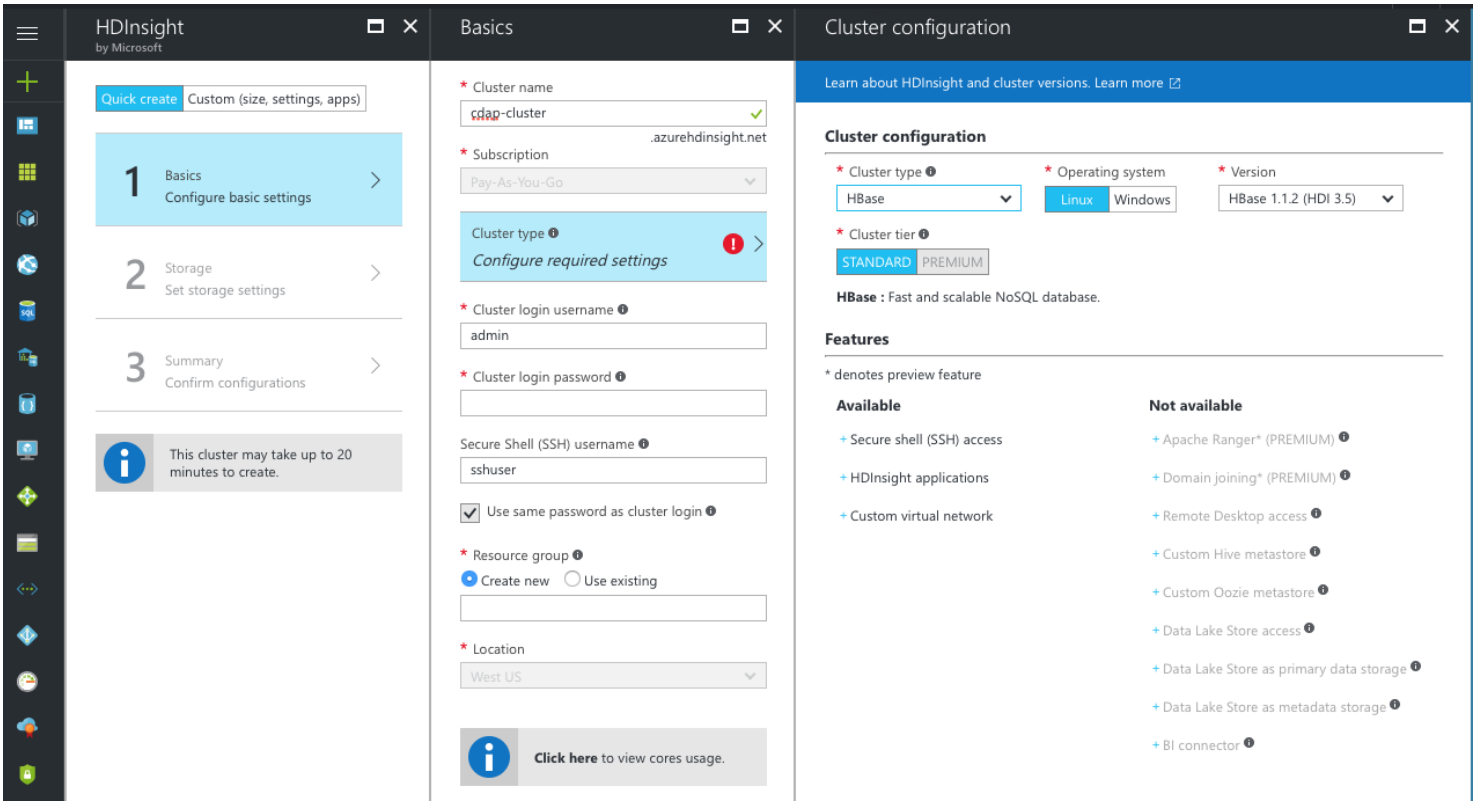
1- Azure HDInsight :

Azure HDInsight est une service cloud des composants Hadoop de Hortonworks Data Platform (HDP). Azure HDInsight permet de traiter des quantités énormes de données facilement, rapidement et à moindre coût. Vous pouvez utiliser les frameworks open source les plus populaires tels que Hadoop, Spark, Hive, LLAP, Kafka, Storm, R.

Le processus de traitement de la donnée avec Spark Streaming :



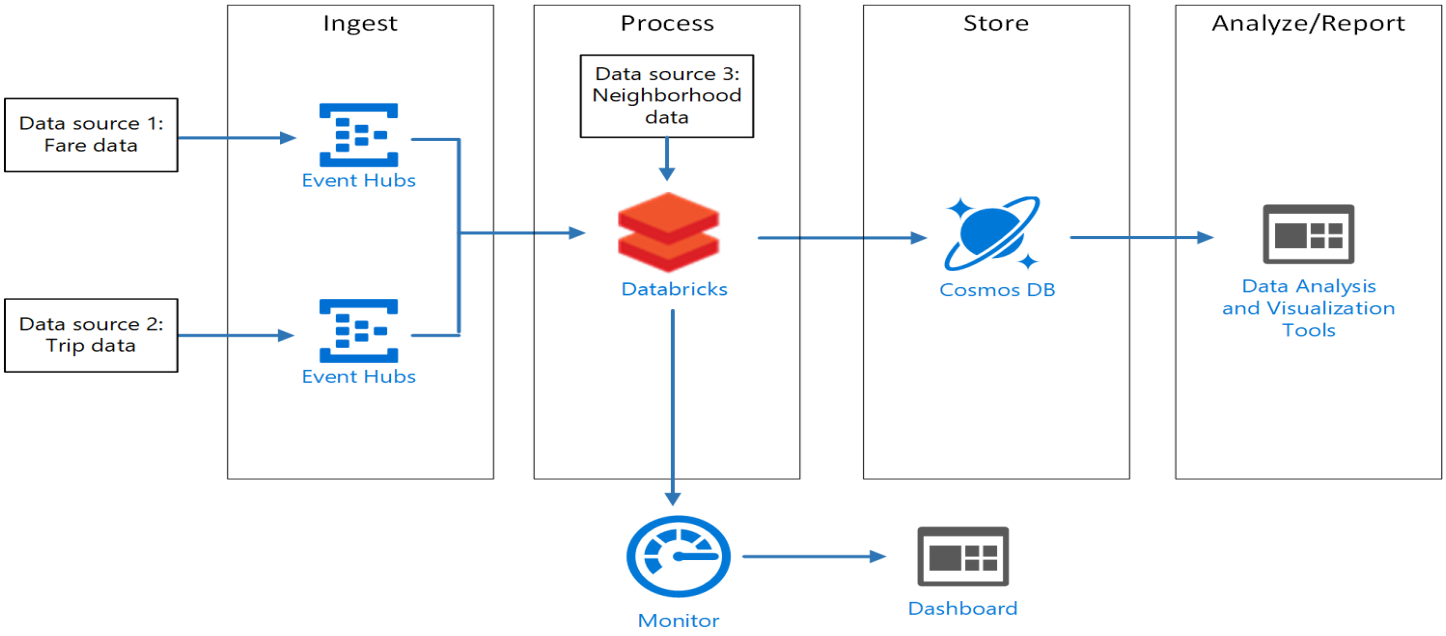
Création et configuration d’un cluster HDInsight dans Microsoft Azure :



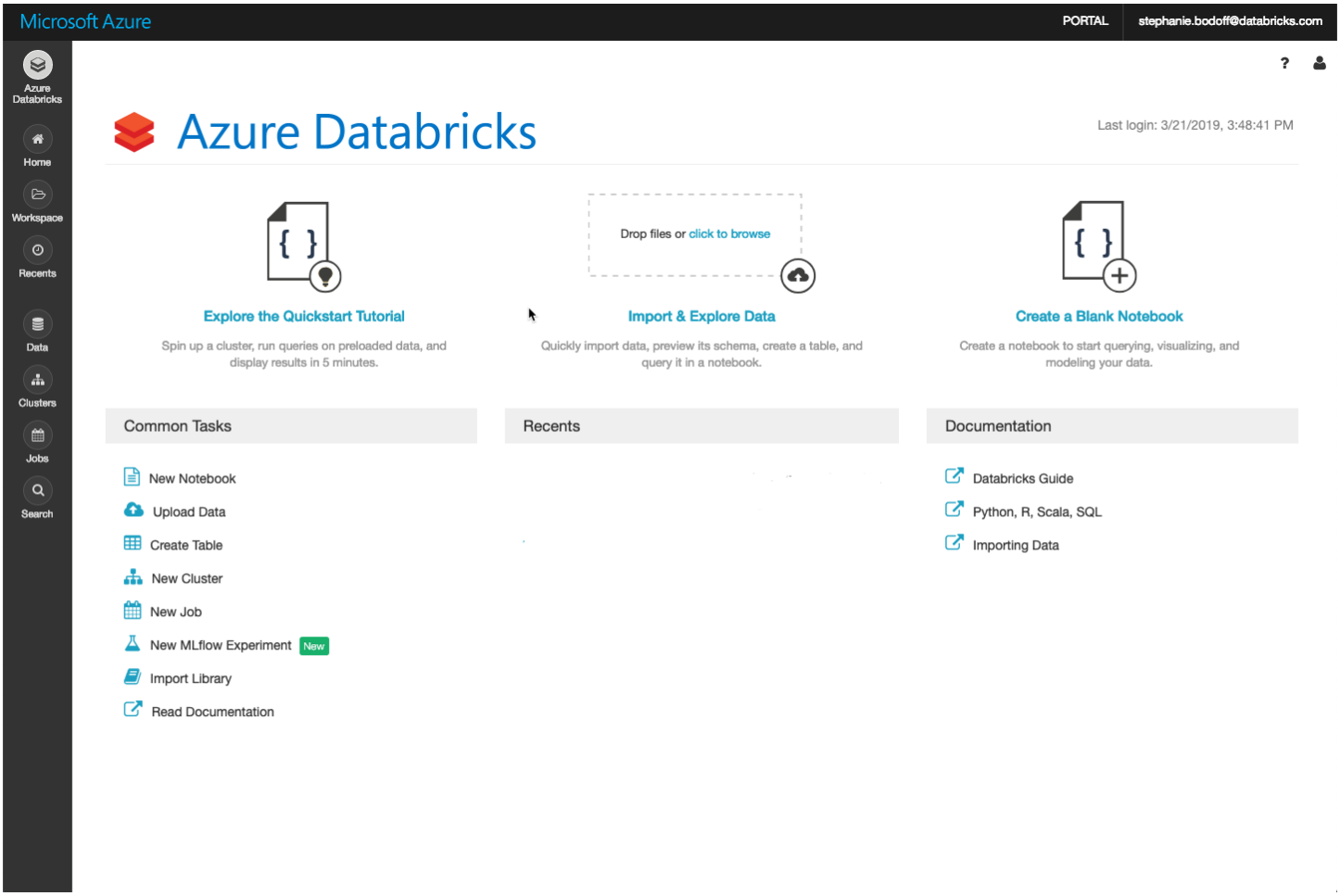
2- Azure Databricks :

Azure Databricks est une plateforme d'analyse basée sur Apache Spark.

Le processus d’ingestion et de traitement de la donnée avec Azure Databricks :



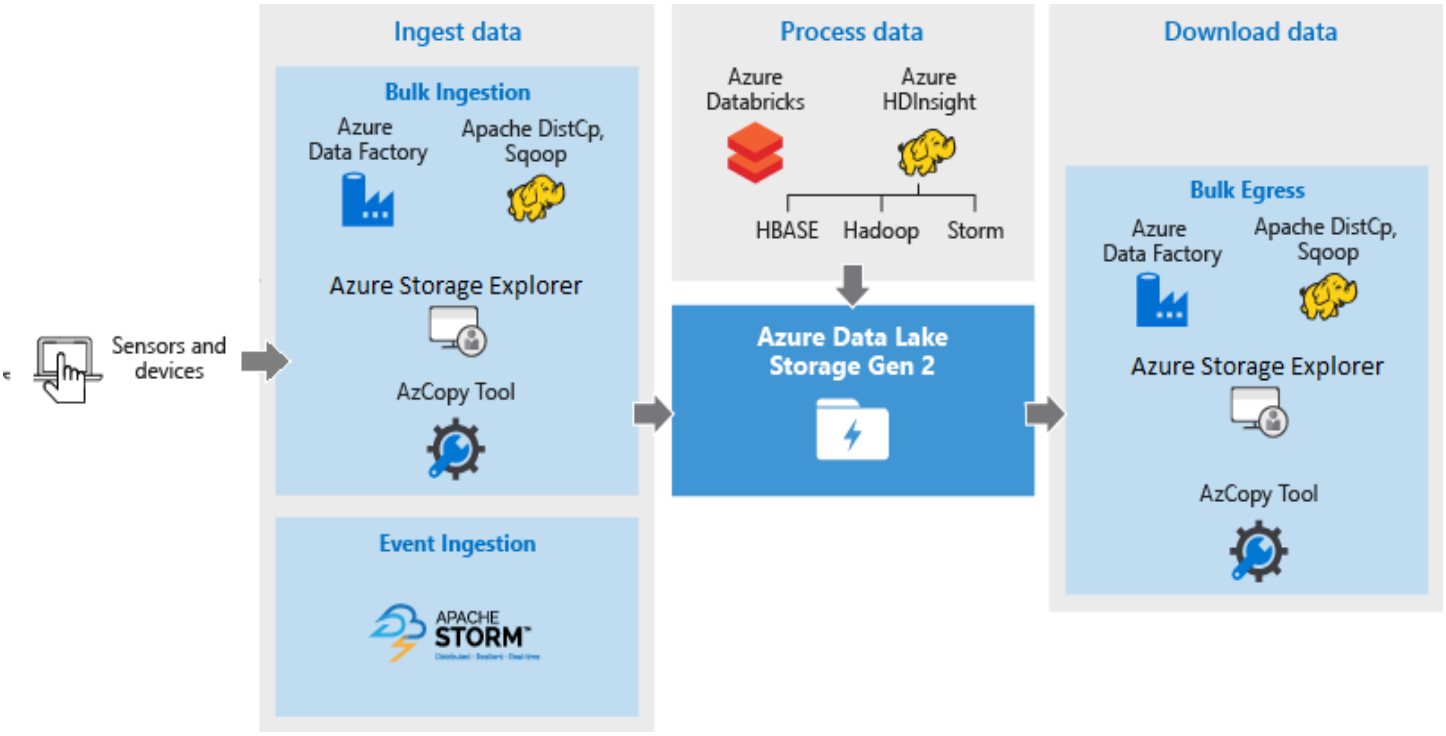
Databricks est optimisé par Apache Spark et offre un panel d'API pouvant être utilisée par les langages R, SQL, Python, Scala et Java. L'écosystème Spark permet également de faire du Streaming avec MLib et GraphX.



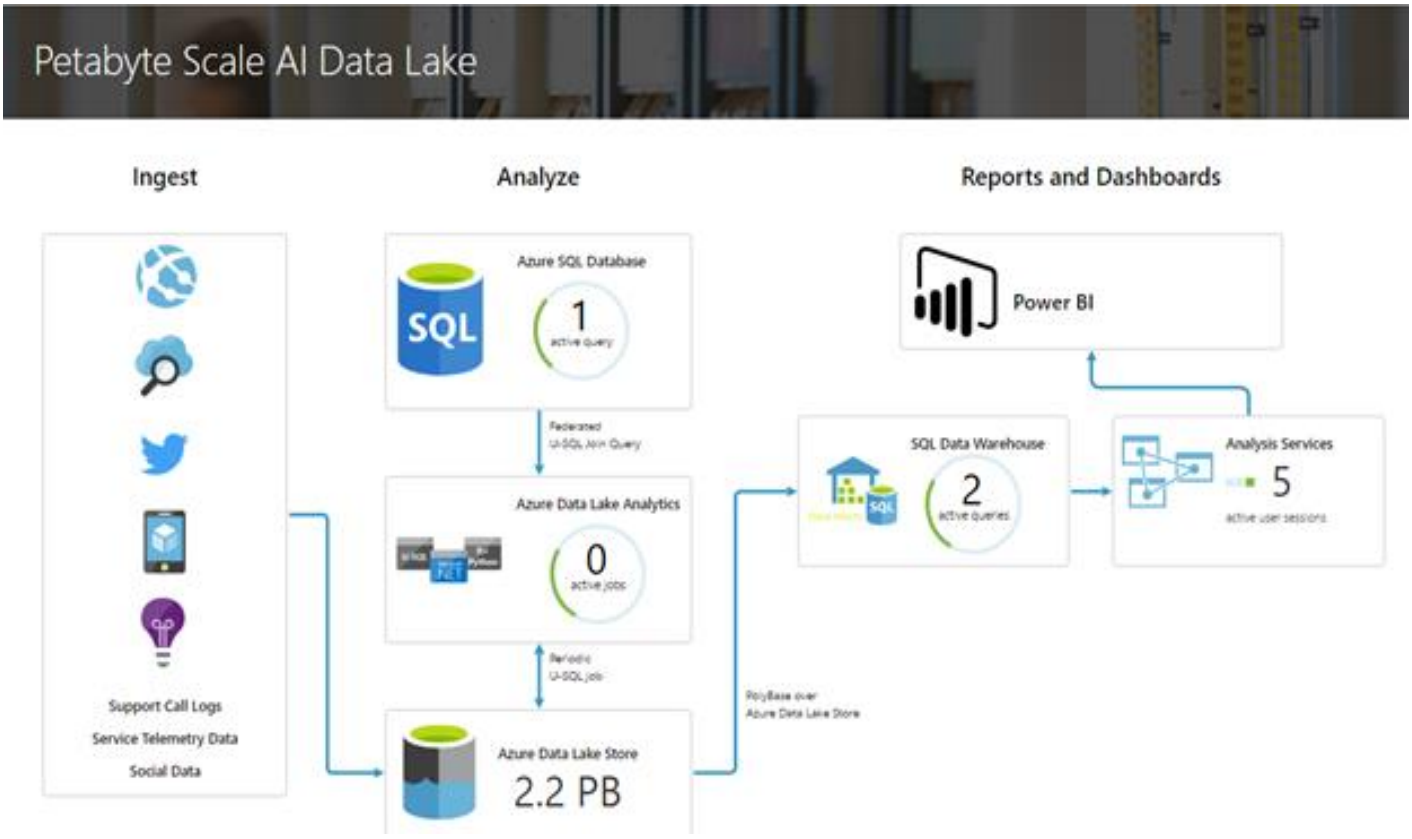
3- Azure Data Lake Analytics :

Azure Data Lake Analytics est un service managé d'analyse à la demande qui simplifie le Big Data. Au lieu de déployer, configurer et ajuster le matériel, vous écrivez des requêtes pour transformer vos données et extraire des informations précieuses.

Le processus d'ingestion et de traitement de la donnée avec Azure Data Lake Storage Gen2 :



Data Lake Analytics offre des fonctionnalités similaires à Databricks. Vous pouvez écrire du code pour analyser les données et l'analyse peut être automatiquement parallélisée. Les données stockées dans un Data Lake sont accessibles de la même manière que sur un volume HDFS.



V- Architectures du big data avec IBM Cloud :

IBM, en partenariat avec Cloudera, fournit la plateforme et les solutions analytiques nécessaires pour créer, gouverner, gérer et explorer votre lac de données Hadoop. ainsi qu'un écosystème de produits et services intégrés.

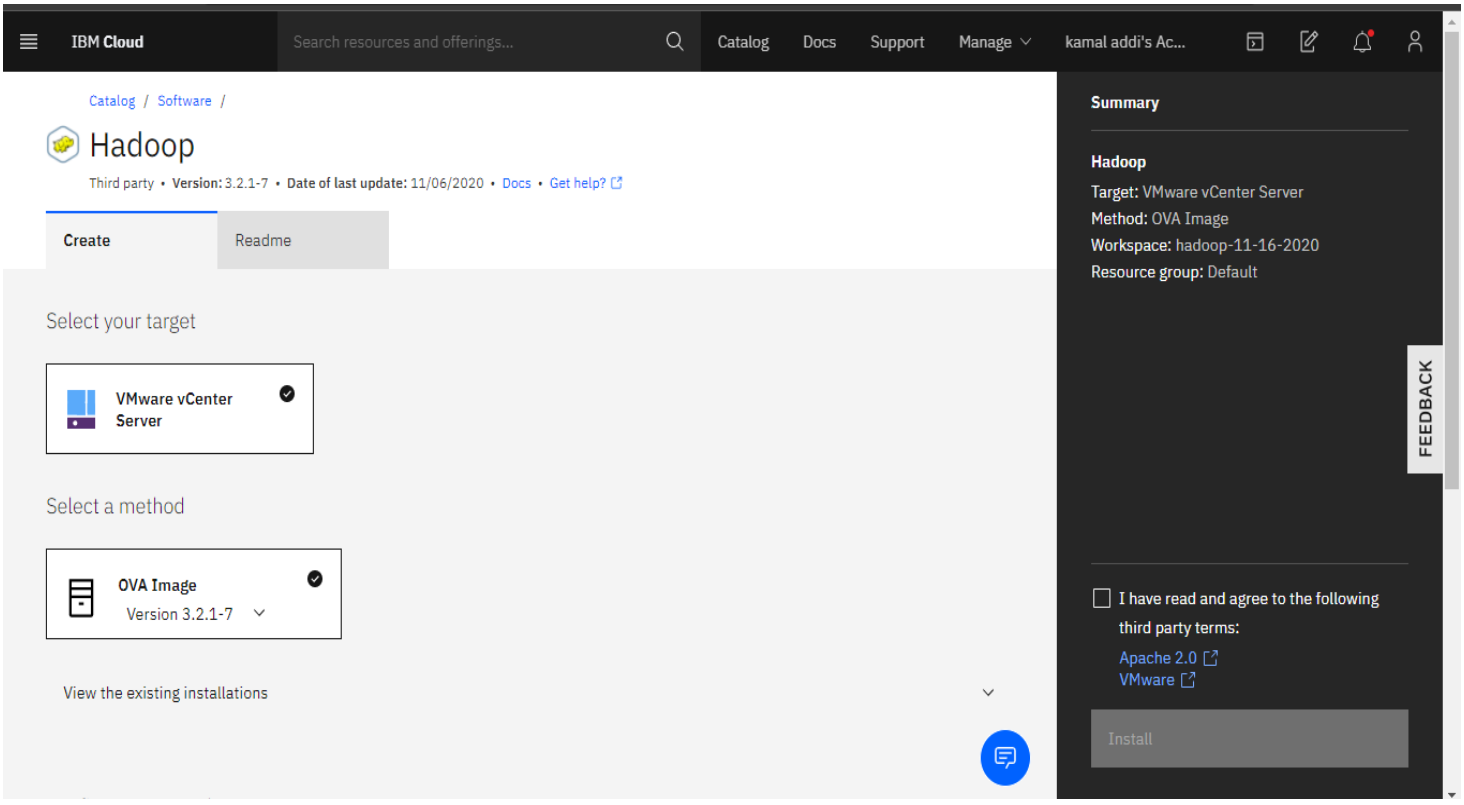
IBM fournit :

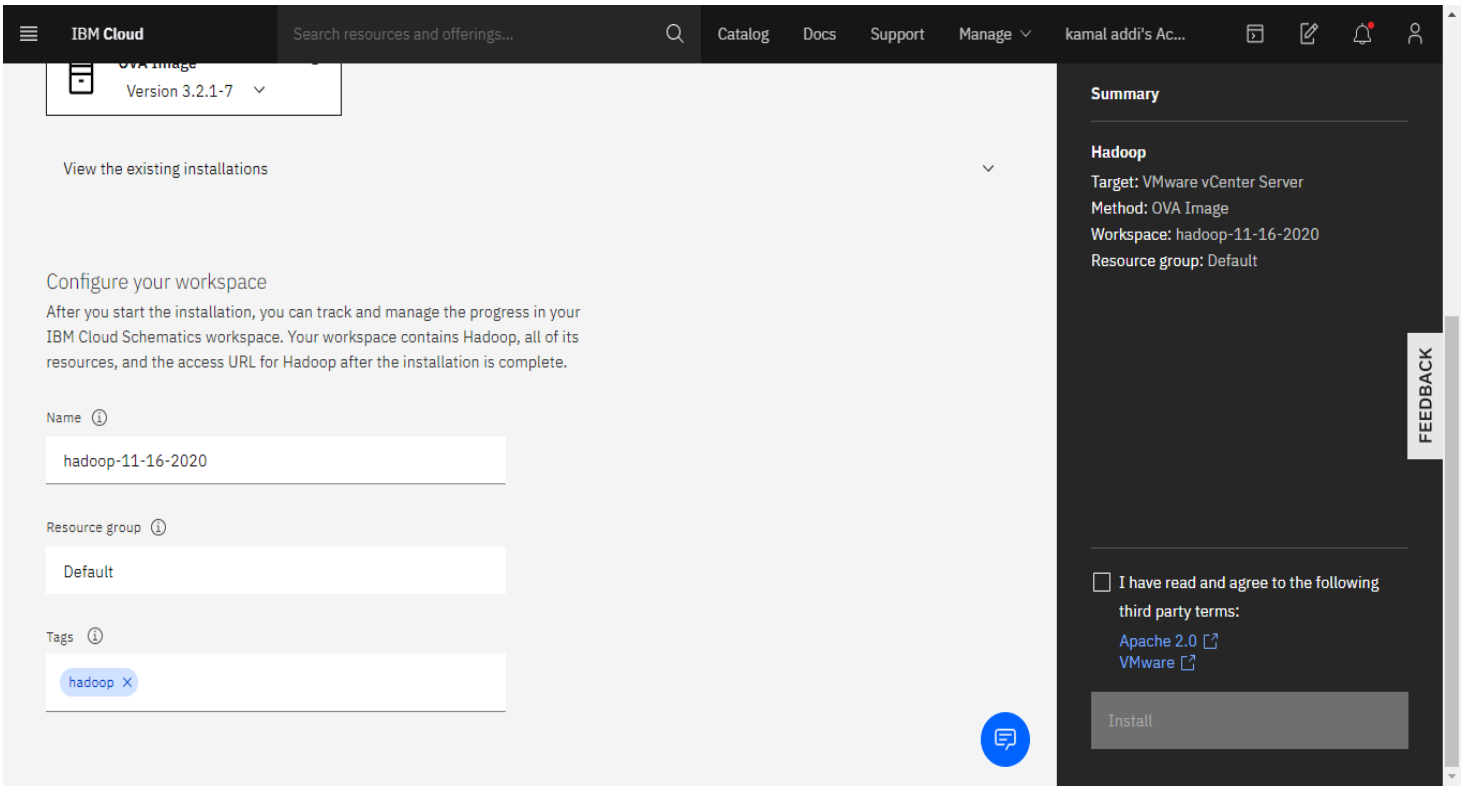
- La revente et le support des produits Cloudera.
- La vente et le support des produits Hortonworks dans le cadre d'un contrat pluriannuel.
- L'aide à la migration vers les futurs produits Cloudera/Hortonworks.

En quelques clics, vous pouvez faire tourner un cluster hadoop multi-nœuds qui est provisionné en quelques minutes. Il est basé sur IBM Open Platform avec Apache Spark et Apache Hadoop. Vous pouvez exécuter des tâches Spark.

Création d'un cluster Hadoop avec IBM Cloud en quelques minutes :

Dans l'onglet de recherche on tape Hadoop, puis la fenêtre ci desous s'ouvre, on choisit la version et le nom de notre cluster et enfin on clique sur install pour installer le framework.





Parmi les avantages d'IBM Cloud aussi qu'il offre la possibilité de déplacer des données entre HDFS et ObjectStore, vous pouvez désormais sauvegarder les données HDFS dans Object Storage.