

ECOSYSTEMS DU BIG DATA DANS LES PLATEFORMES 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 :
- II- Architectures du big data avec Google Cloud Platform (GCP) :
- III- Architectures du big data avec AWS :
- IV- Architectures du big data avec Microsoft Azure :
- V- Architectures du big data avec IBM Cloud :

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 peuvent expérimenter ces besoins via des plateformes tels que Google Cloud Plateforme, 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 un 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 :

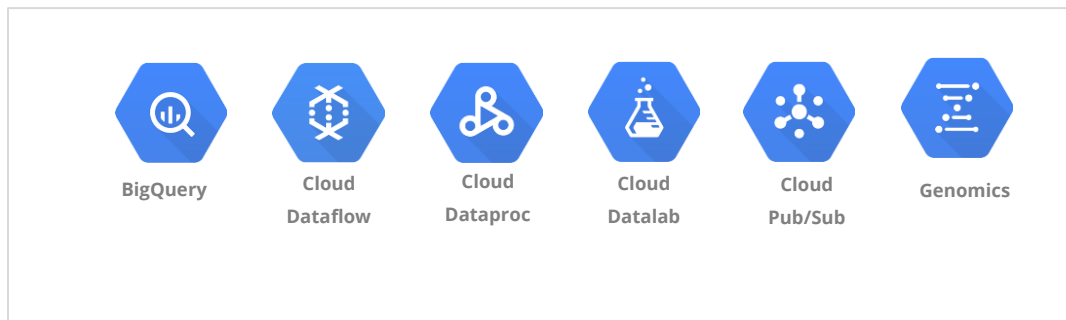


Fig1 : *Google Cloud Platform products and services for big data.*

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.

Big Data Lifecycle

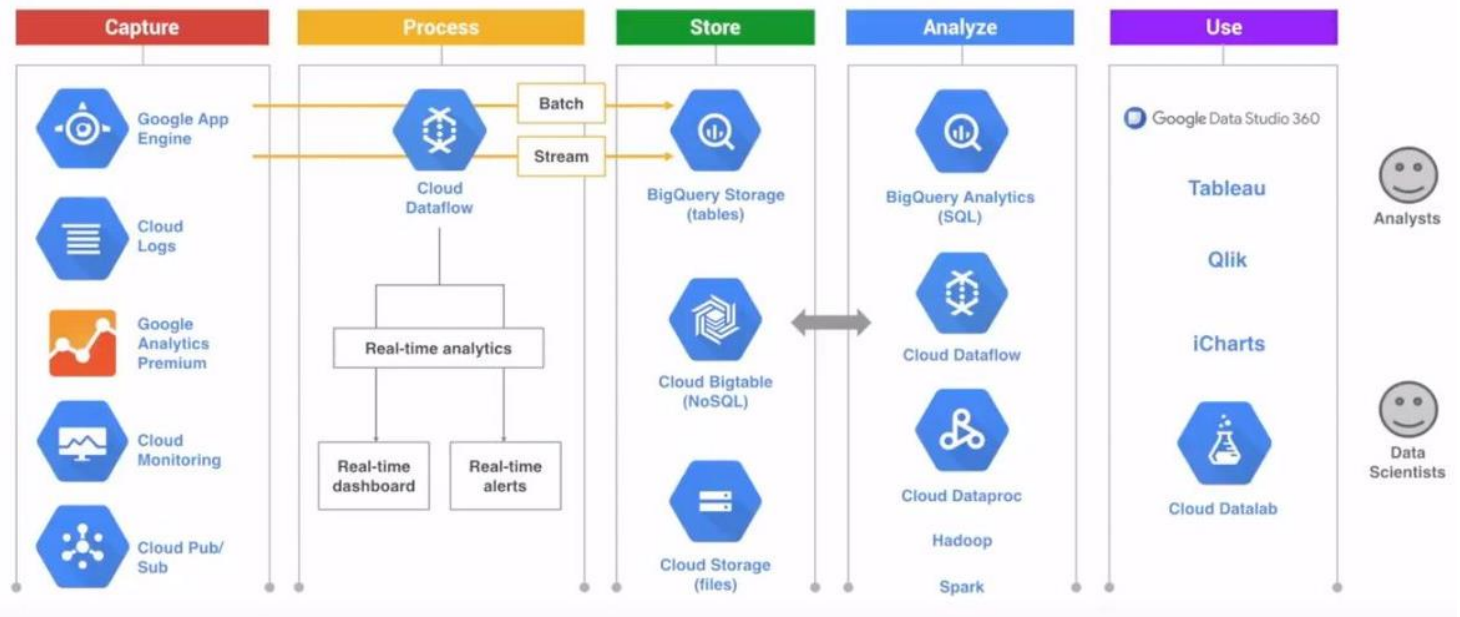
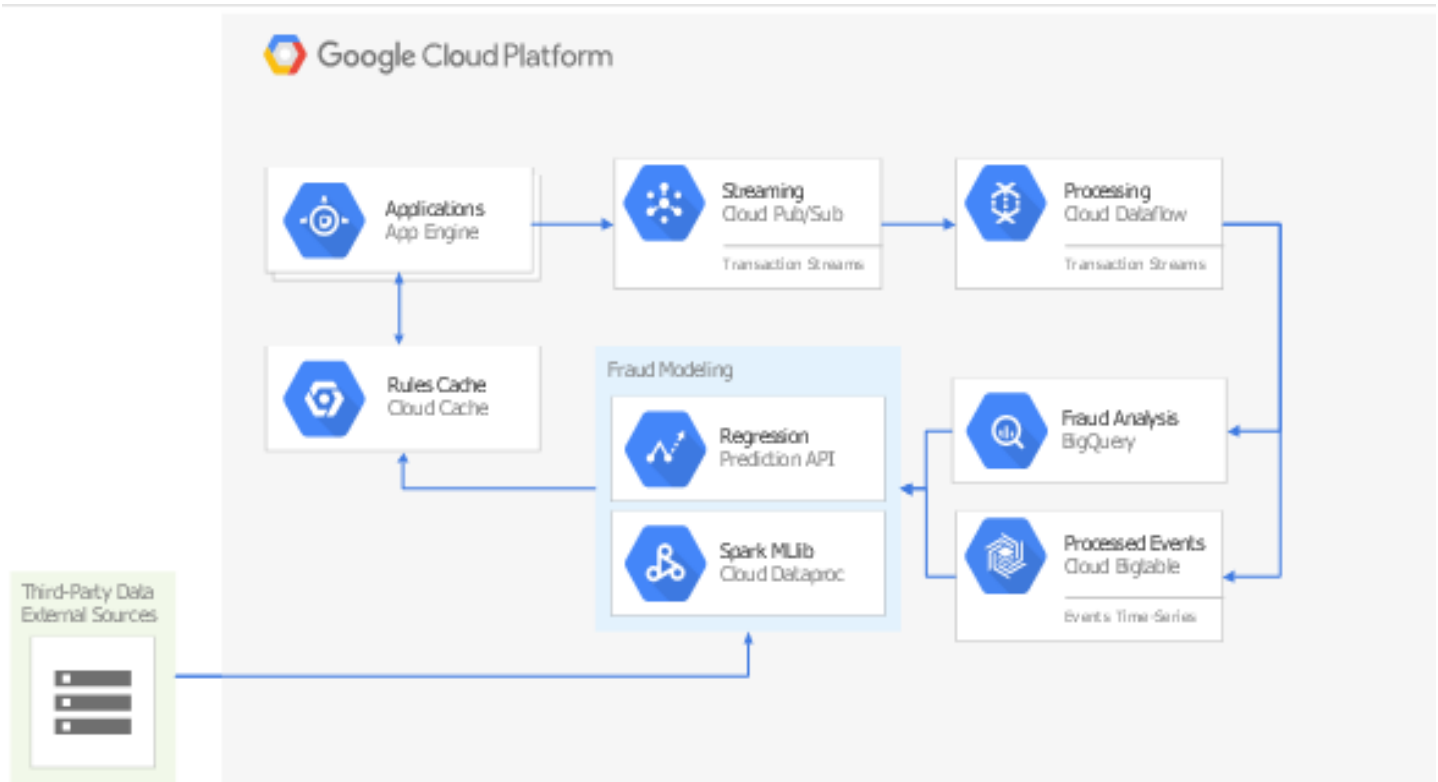
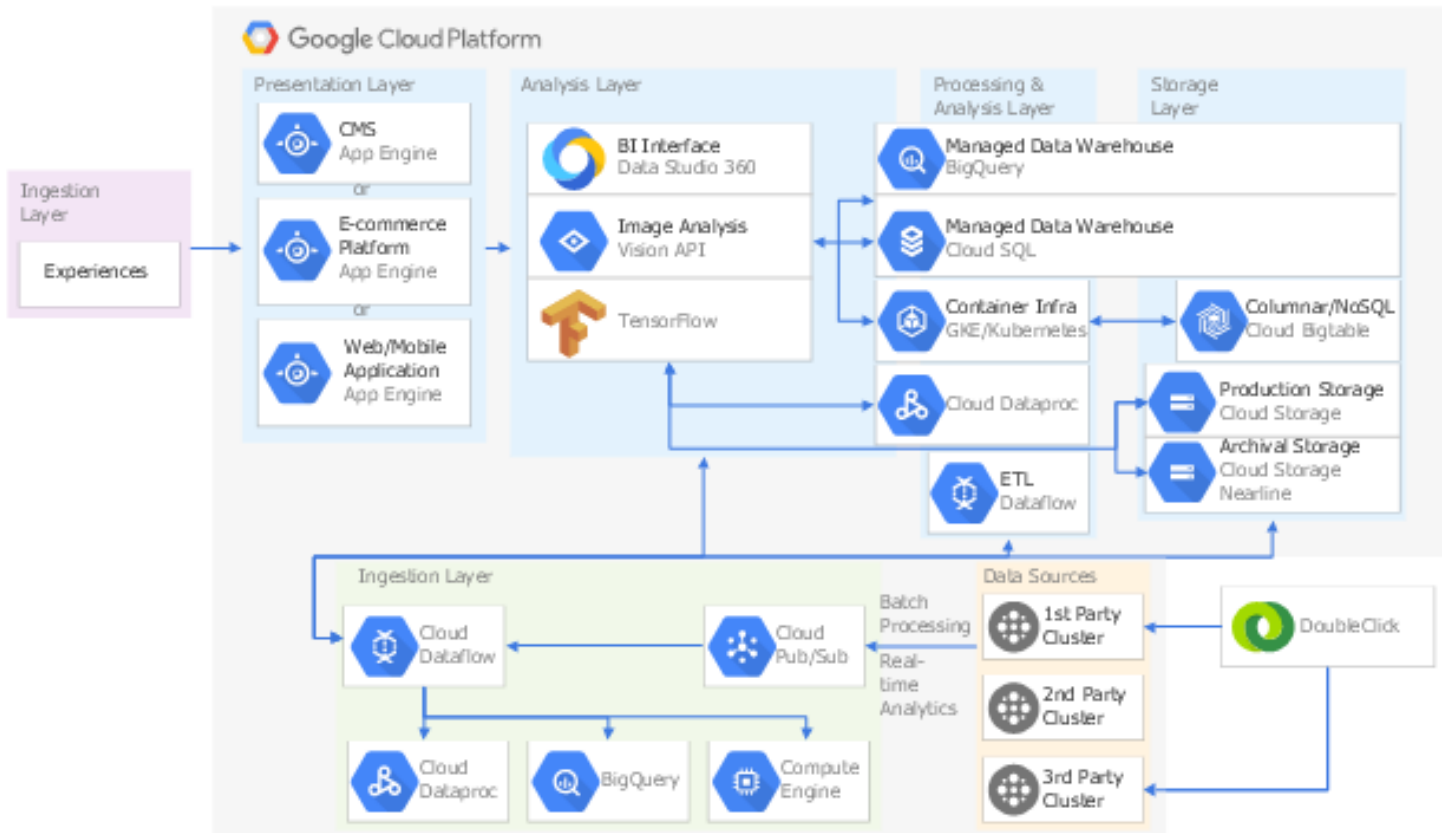


Fig2 : le cycle de vie d'une architecture big data.

1- Detection ds fraudes :

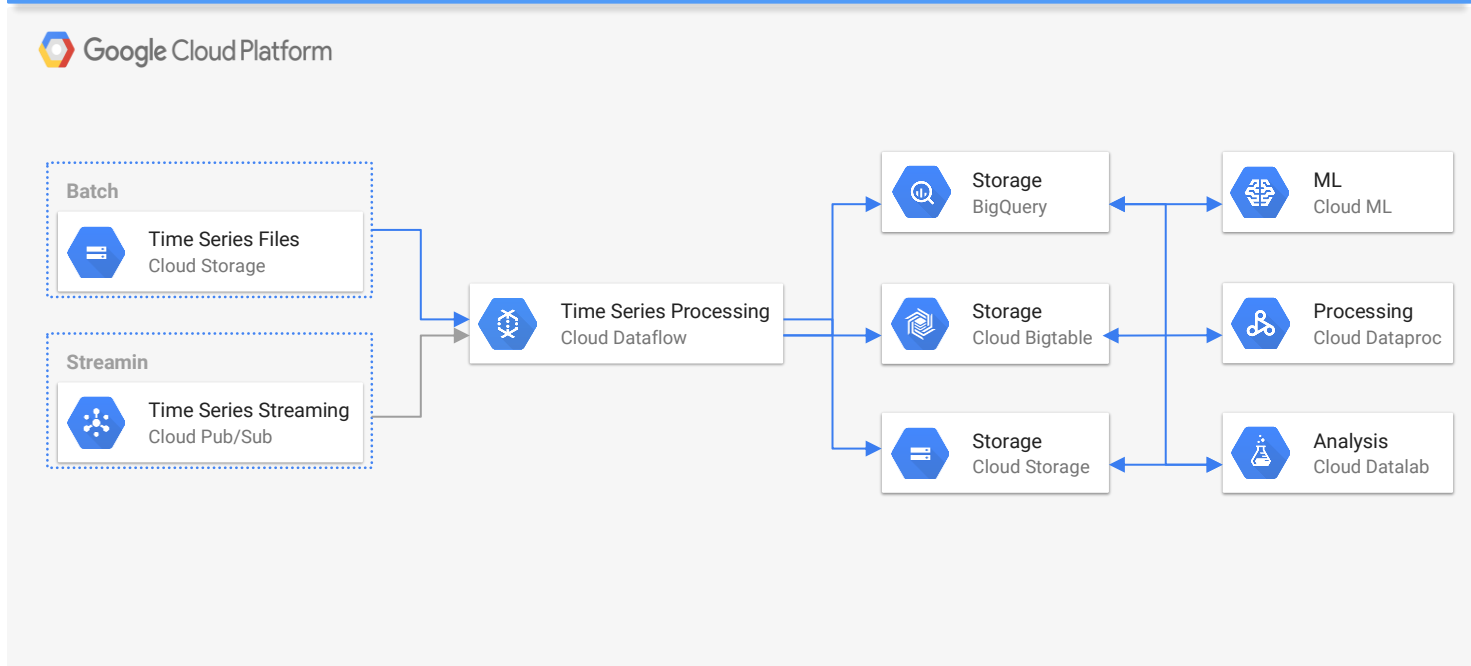


2- Digital marketing :



3- Services financiers et analyse de séries temporelles :

Architecture: Big Data > Time Series Analysis

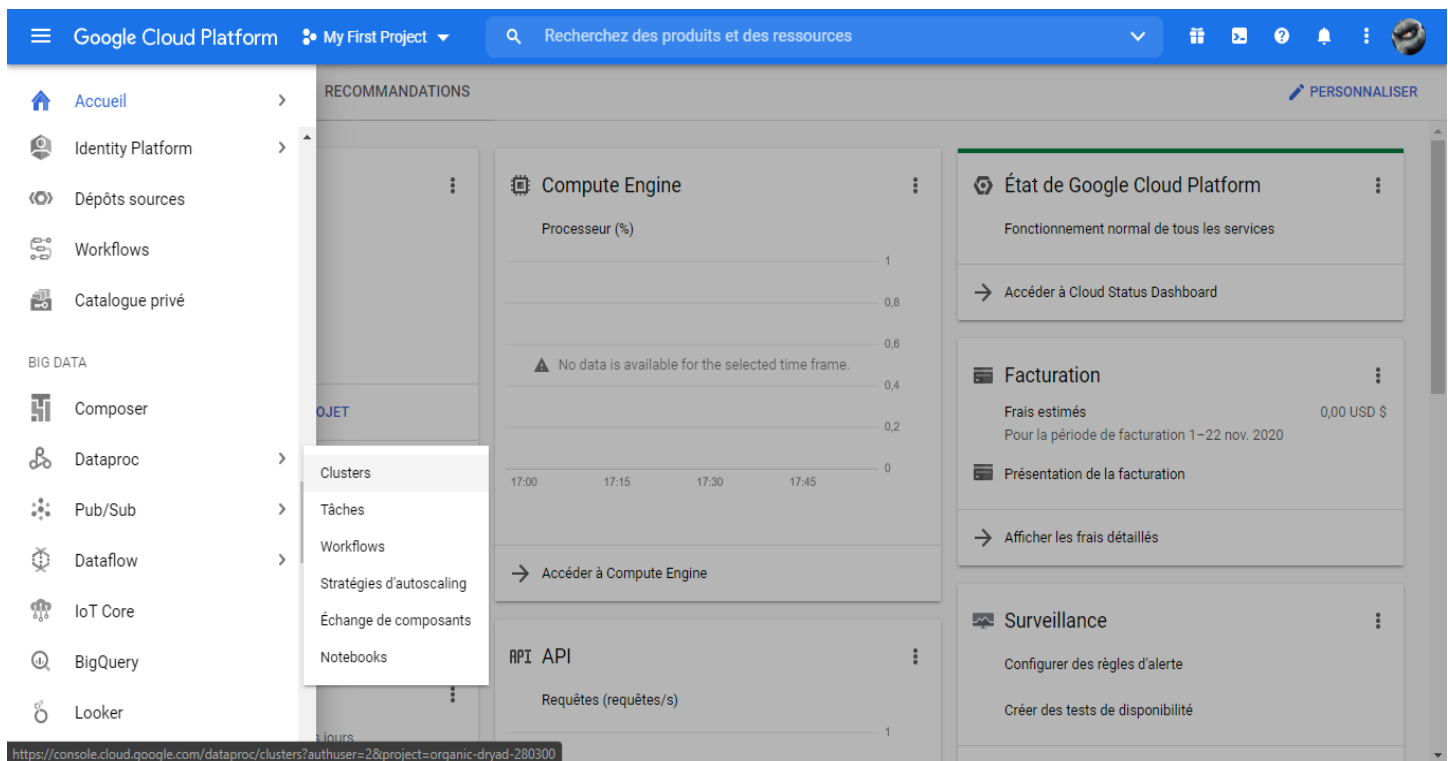


Dataprocc :

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

Cloud Dataprocc 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 Dataprocc : sélectionné l'onglet dataprocc puis clusters :



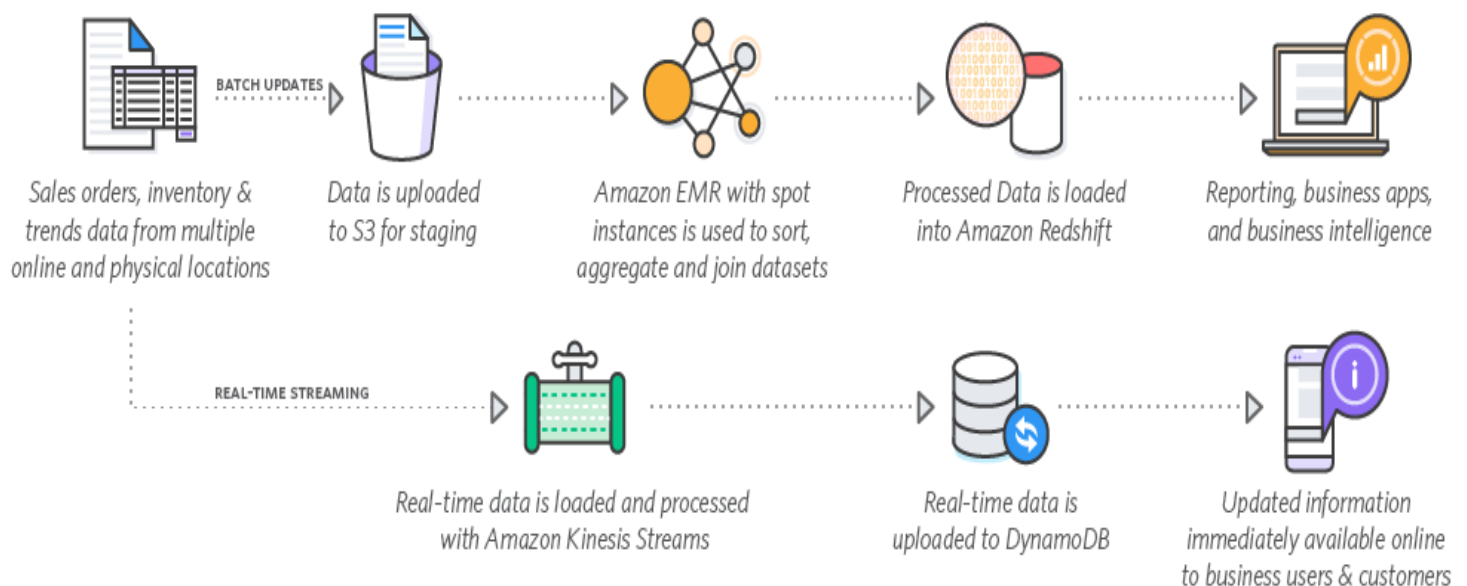
Une fenêtre de configuration s'ouvre sur lequel on choisit le nombre de neuds, memoires, CPU etc..

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.

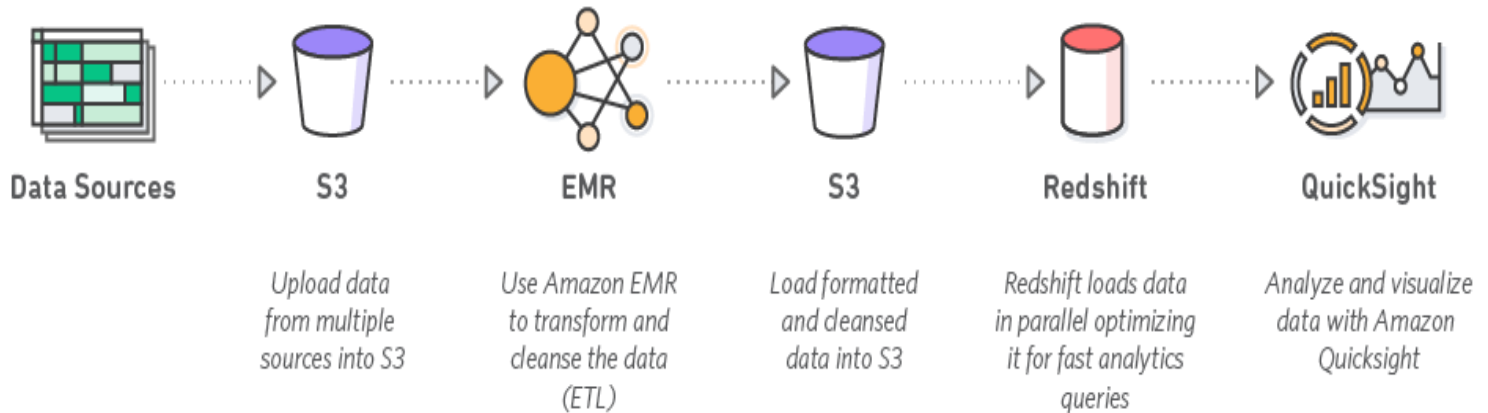
1- Analyses big data à la demande :

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.



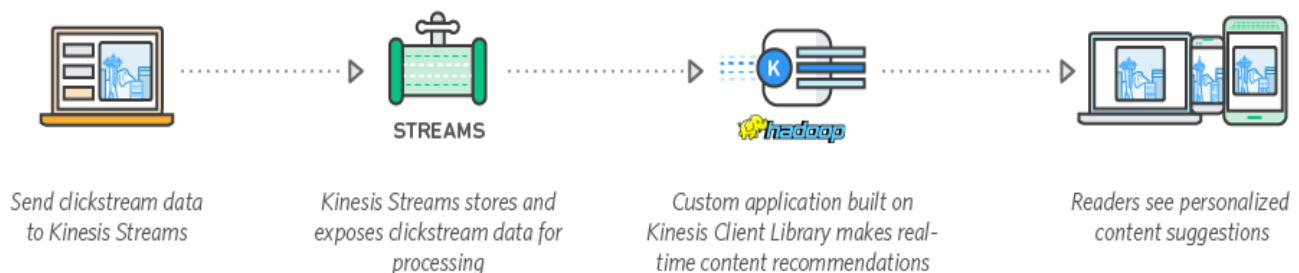
2- Entreposage de données :

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 :

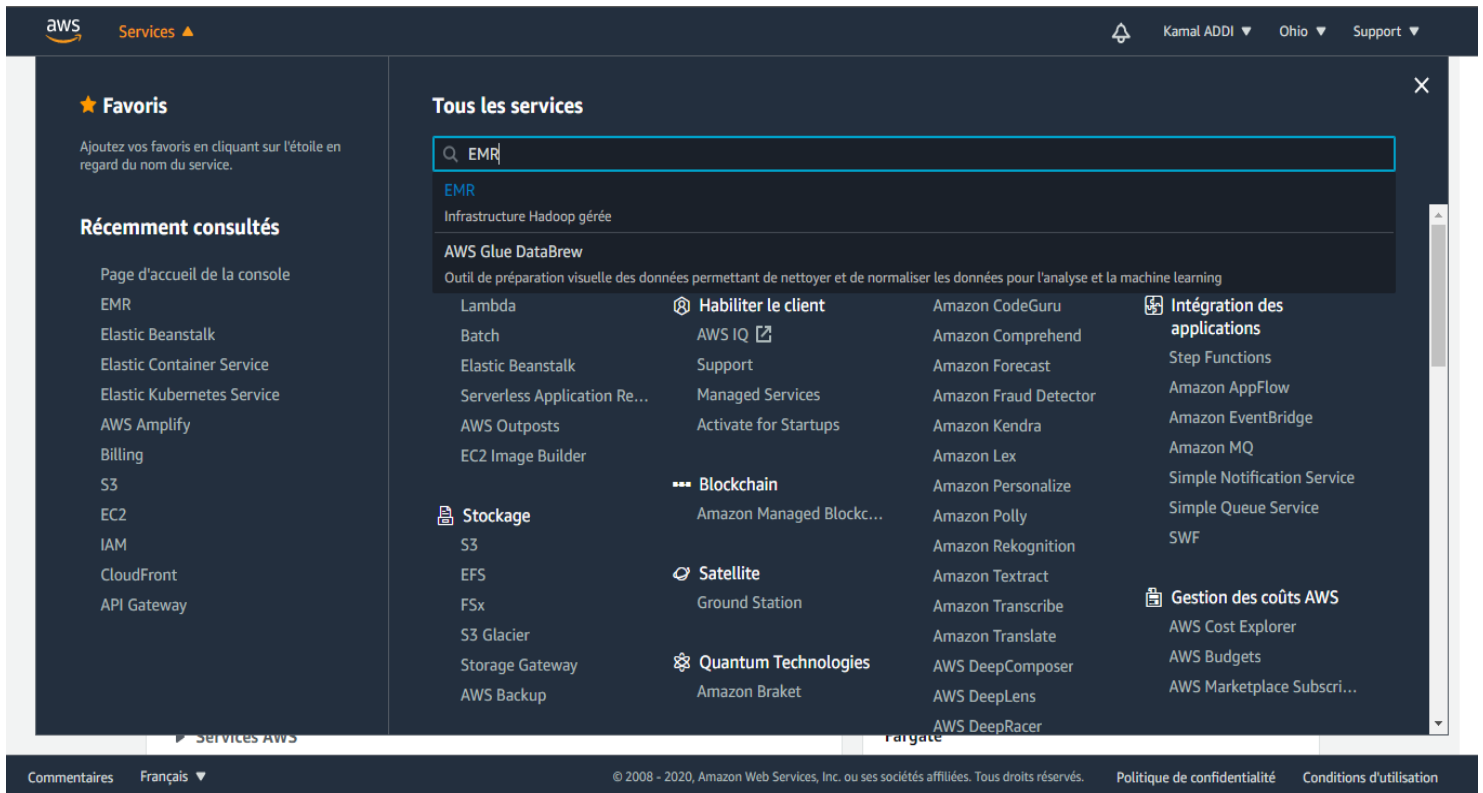
Améliorez l'expérience numérique de vos clients et apprenez à mieux connaître votre site Web. Collectez, traitez, analysez et visualisez des analyses de parcours de navigation en temps réel avec AWS.



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

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

The screenshot shows the AWS EMR console interface for creating a new cluster. The top navigation bar includes the AWS logo, 'Services', and user information (Kamal ADDI, Ohio, Support). The main heading is 'Créer un cluster - Options rapides' with a link to 'Accéder aux options avancées'.

Configuration générale

- Nom du cluster:** Mon cluster
- Journalisation:** ☒ (Information icon)
- Dossier S3:** s3://aws-logs-532299495469-us-east-2/elasticmapred (Folder icon)
- Mode de lancement:** ☒ Cluster (Information icon) ☐ Exécution d'étape (Information icon)

Configuration des logiciels

- Libérer:** emr-5.31.0 (Information icon)
- Applications:**
 - ☒ Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
 - ☐ HBase: HBase 1.4.13, Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and ZooKeeper 3.4.14
 - ☐ Presto: Presto 0.238.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore
 - ☐ Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2
 - ☐ Utiliser AWS Glue Data Catalog pour les métadonnées de table (Information icon)

Configuration du matériel

- Type d'instance:** m5.xlarge (Information icon). Note: Le type d'instance sélectionné ajoute un volume EBS GP2 par défaut de 64 GiO par instance. [En savoir plus](#)
- Nombre d'instances:** 3 (1 nœud maître et 2 nœuds principaux)
- Cluster scaling:** ☐ scale cluster nodes based on workload

Sécurité et accès

- Paire de clés EC2:** Choisir une option (Information icon). [Apprenez à créer une paire de clés EC2.](#)
- Autorisations:** ☒ Par défaut ☐ Personnalisé. Note: Utilisez les rôles IAM par défaut. Si des rôles sont absents, ils seront créés automatiquement pour vous avec des stratégies gérées pour les mises à jour automatiques de stratégies.
- Rôle EMR:** EMR_DefaultRole (Information icon)
- Profil d'instance EC2:** EMR_EC2_DefaultRole (Information icon)

Buttons at the bottom: [Annuler](#) and [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 :

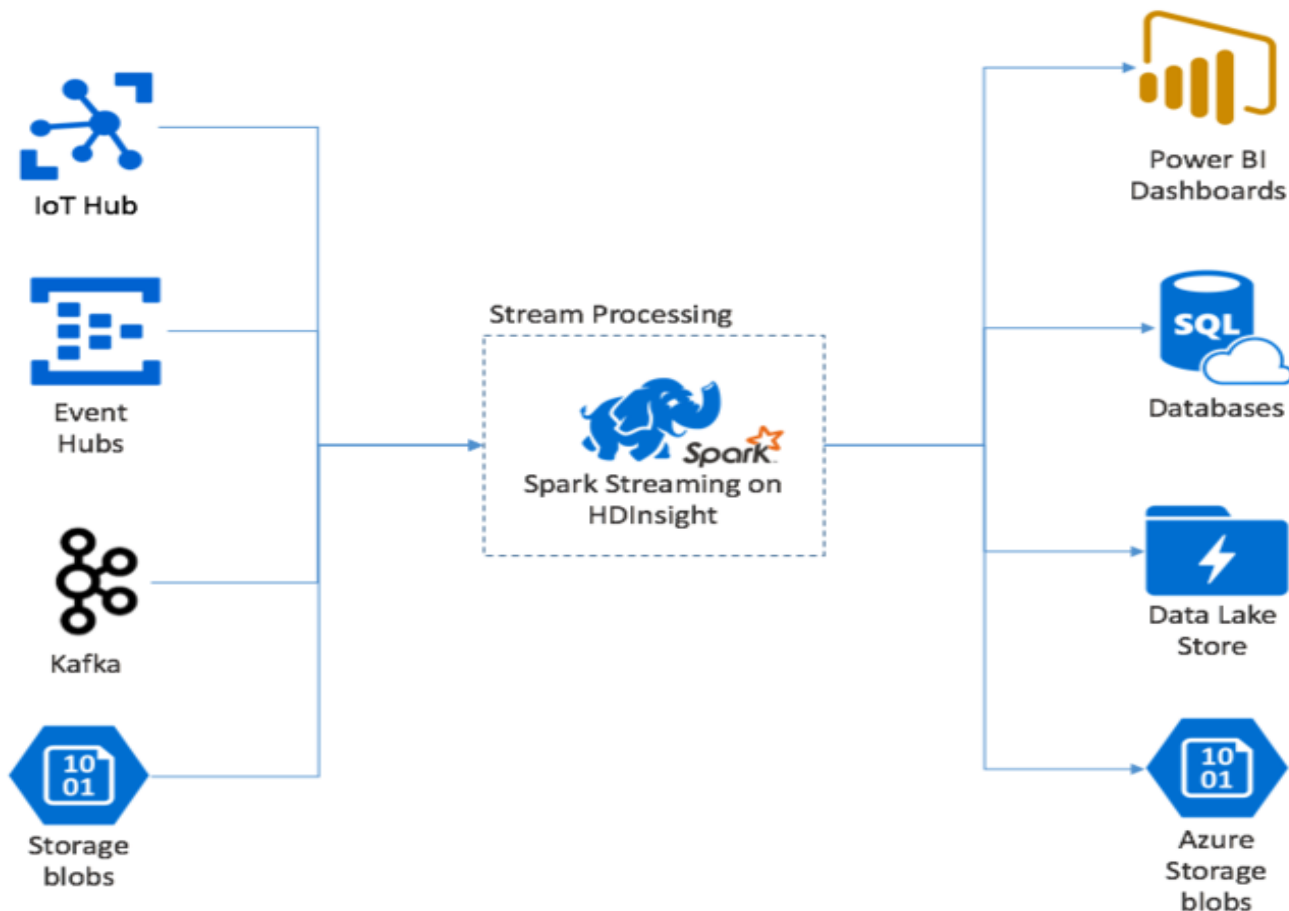
Comme la plupart des plateformes cloud, Microsoft Azure a aussi des outils disponibles sur le marketplace pour la réplication et l'ingestion des données en temps réel.

Les solutions big data proposées par Microsoft sont les suivantes :

1- Azure HDInsight :

Azure HDInsight est un 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 :

HDInsight by Microsoft

Quick create Custom (size, settings, apps)

1 Basics
Configure basic settings

2 Storage
Set storage settings

3 Summary
Confirm configurations

This cluster may take up to 20 minutes to create.

Basics

* Cluster name
hdinsight-cluster

* Subscription
Pay-As-You-Go

Cluster type
Configure required settings

* Cluster login username
admin

* Cluster login password

Secure Shell (SSH) username
sshuser

☒ Use same password as cluster login

* Resource group
☒ Create new ☐ Use existing

* Location
West US

Click here to view cores usage.

Cluster configuration

Learn about HDInsight and cluster versions. Learn more

Cluster configuration

* Cluster type
HBase

* Operating system
Linux Windows

* Version
HBase 1.1.2 (HDI 3.5)

* Cluster tier
STANDARD PREMIUM

HBase : Fast and scalable NoSQL database.

Features

* denotes preview feature

Available

- + Secure shell (SSH) access
- + HDInsight applications
- + Custom virtual network

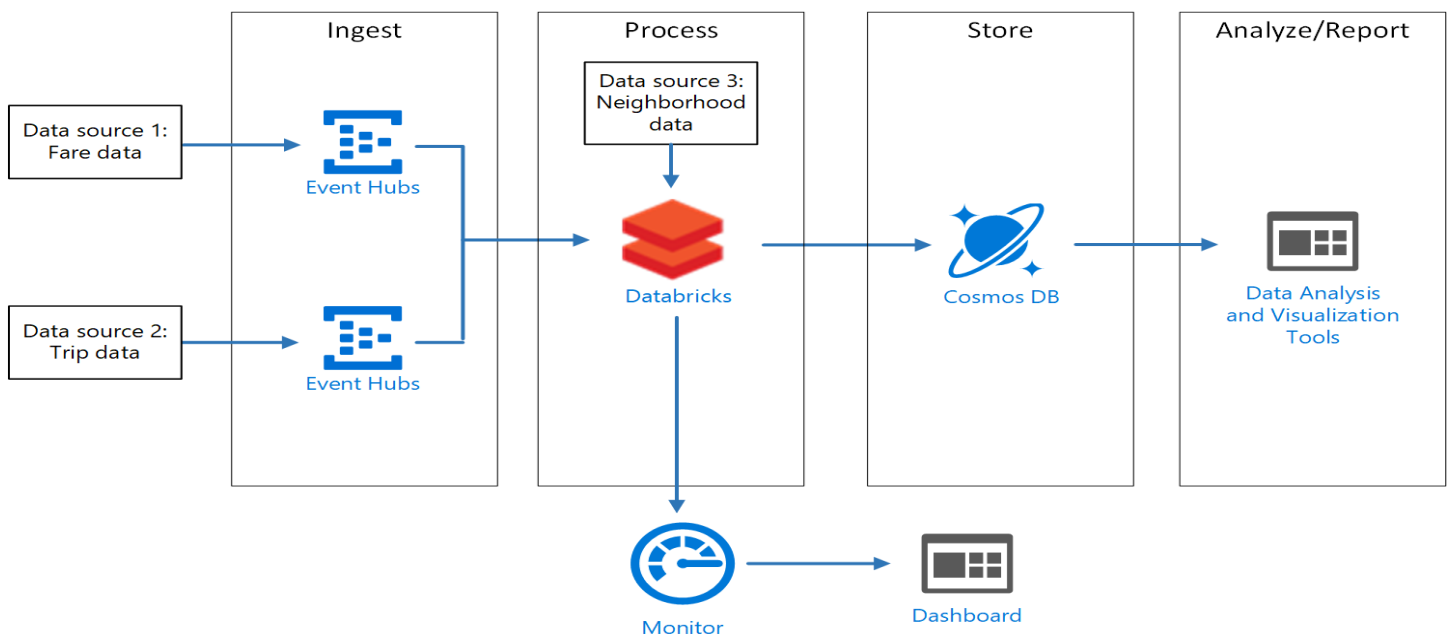
Not available

- + Apache Ranger* (PREMIUM)
- + Domain joining* (PREMIUM)
- + Remote Desktop access
- + Custom Hive metastore
- + Custom Oozie metastore
- + Data Lake Store access
- + Data Lake Store as primary data storage
- + Data Lake Store as metadata storage
- + BI connector

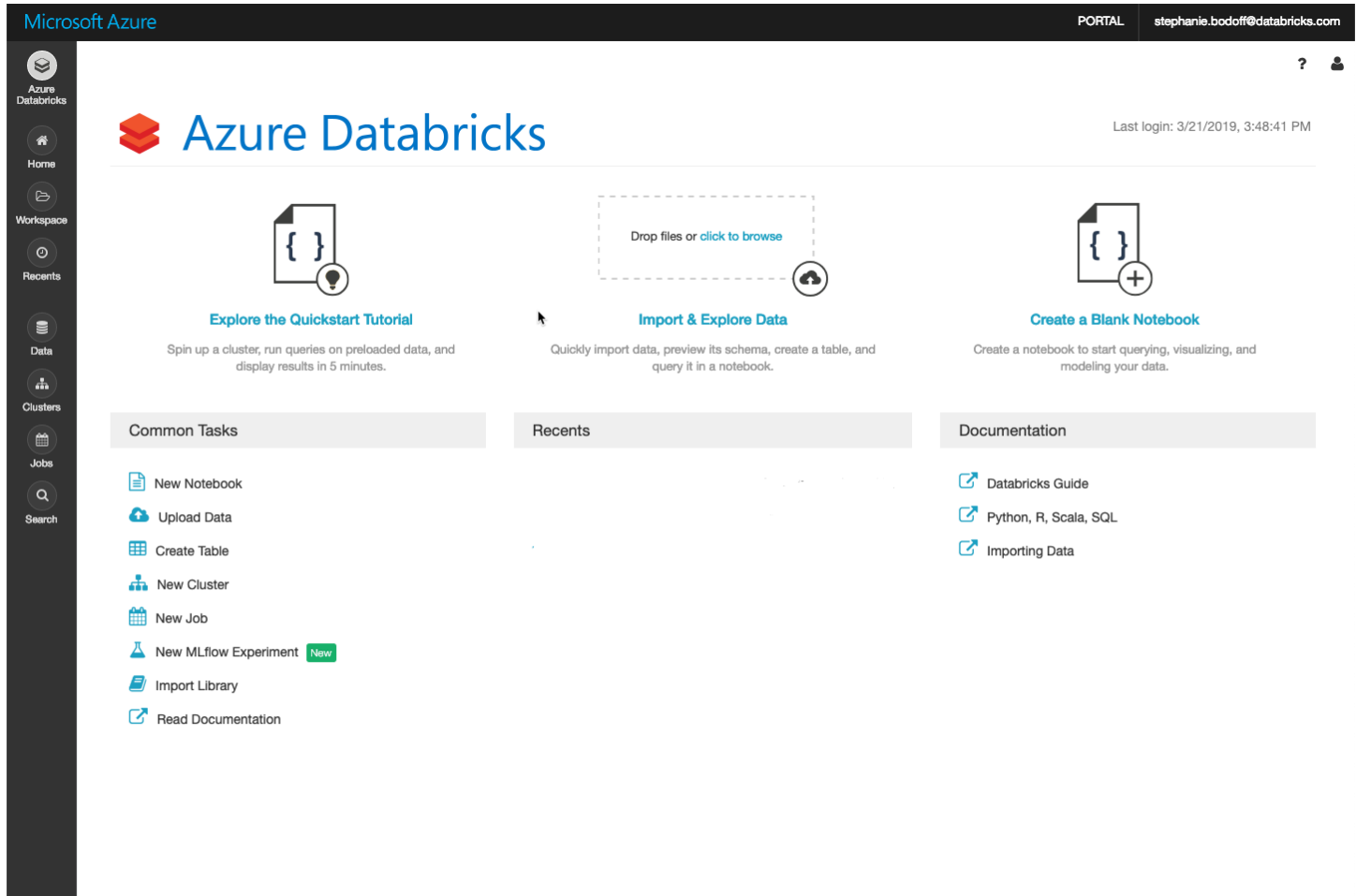
2- Azure Databricks :

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

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



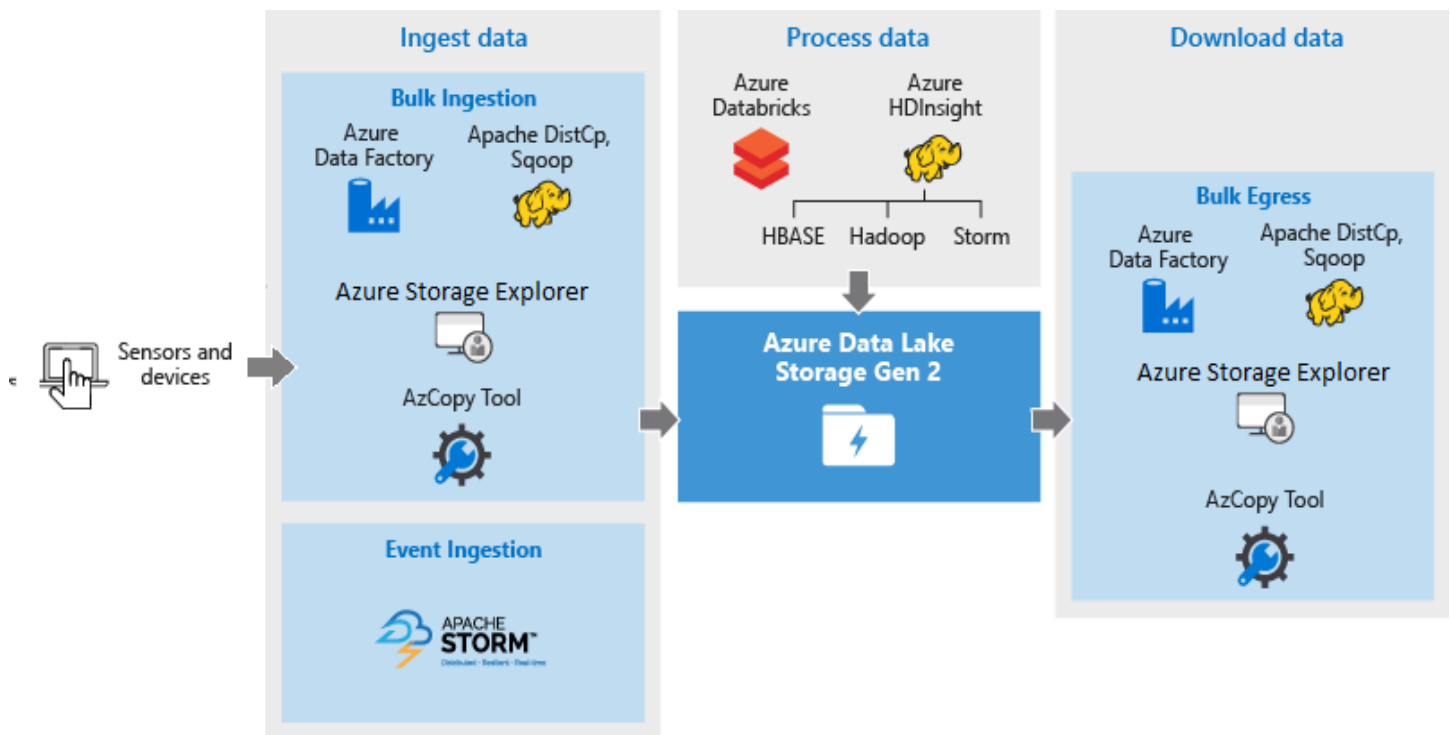
Databricks est optimisé par Apache Spark et offre un panel d'API pouvant être utilisée par les langages R, SQL, Python, Scala et Java. L'écosystème Spark permet également de faire du Streaming avec MLib et GraphX.



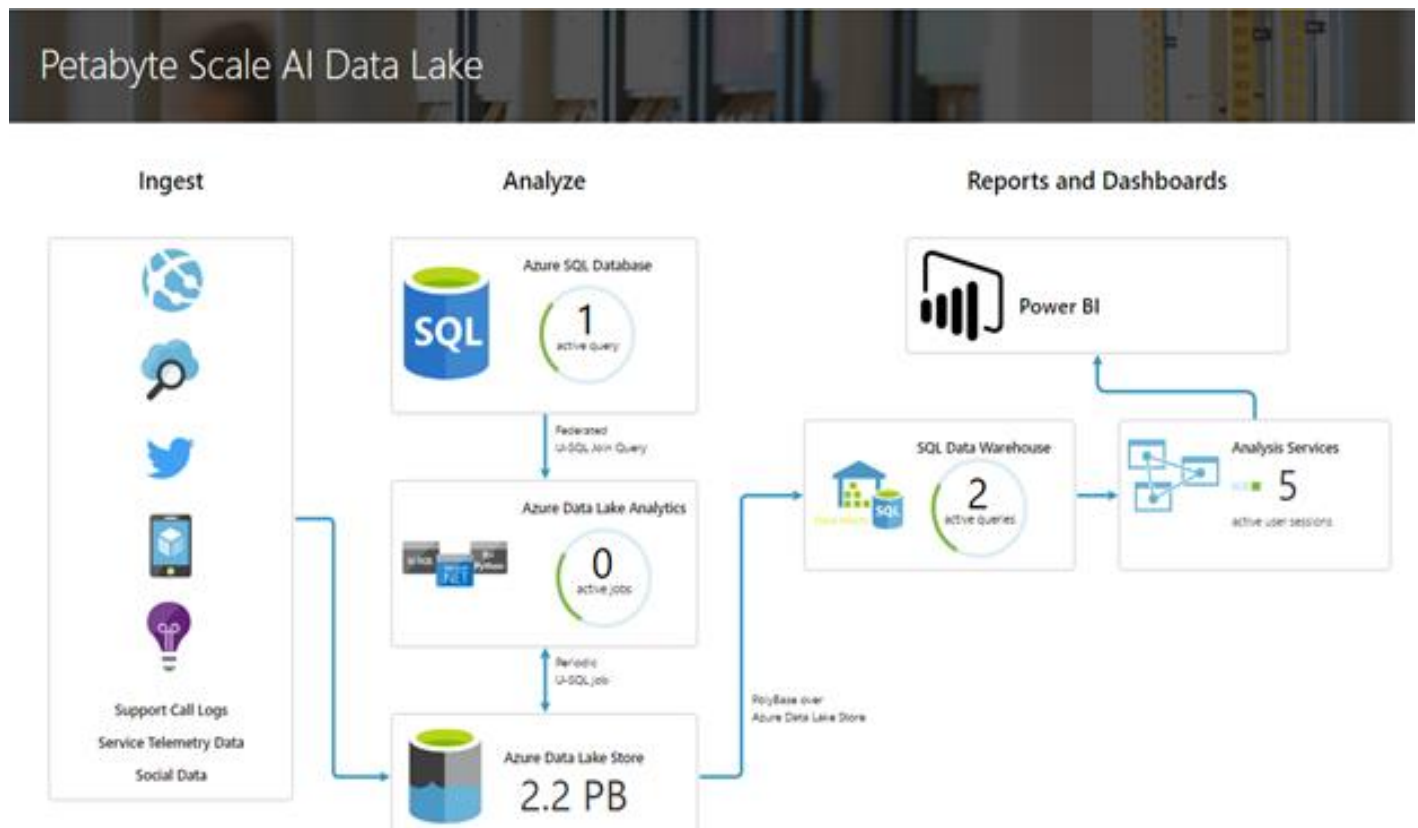
3- Azure Data Lake Analytics :

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.

IBM Cloud

Search resources and offerings...

Catalog Docs Support Manage

kamal addi's Ac...

Catalog / Software /

Hadoop

Third party • Version: 3.2.1-7 • Date of last update: 11/06/2020 • Docs • Get help?

Create Readme

Select your target

VMware vCenter Server

Select a method

OVA Image
Version 3.2.1-7

View the existing installations

Summary

Hadoop

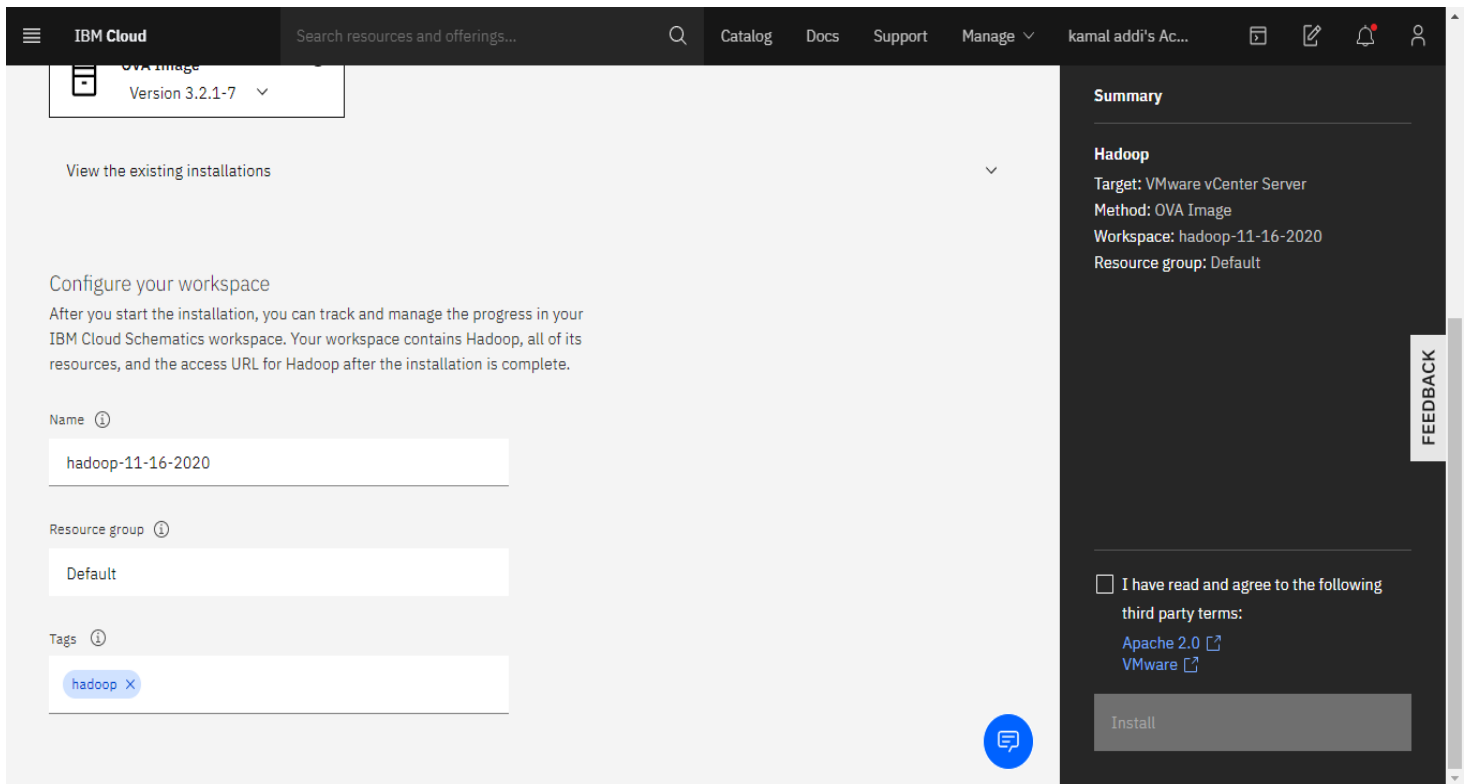
Target: VMware vCenter Server
Method: OVA Image
Workspace: hadoop-11-16-2020
Resource group: Default

☐ I have read and agree to the following third party terms:

Apache 2.0
VMware

Install

FEEDBACK



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.