COLLÈGE ROSEMONT

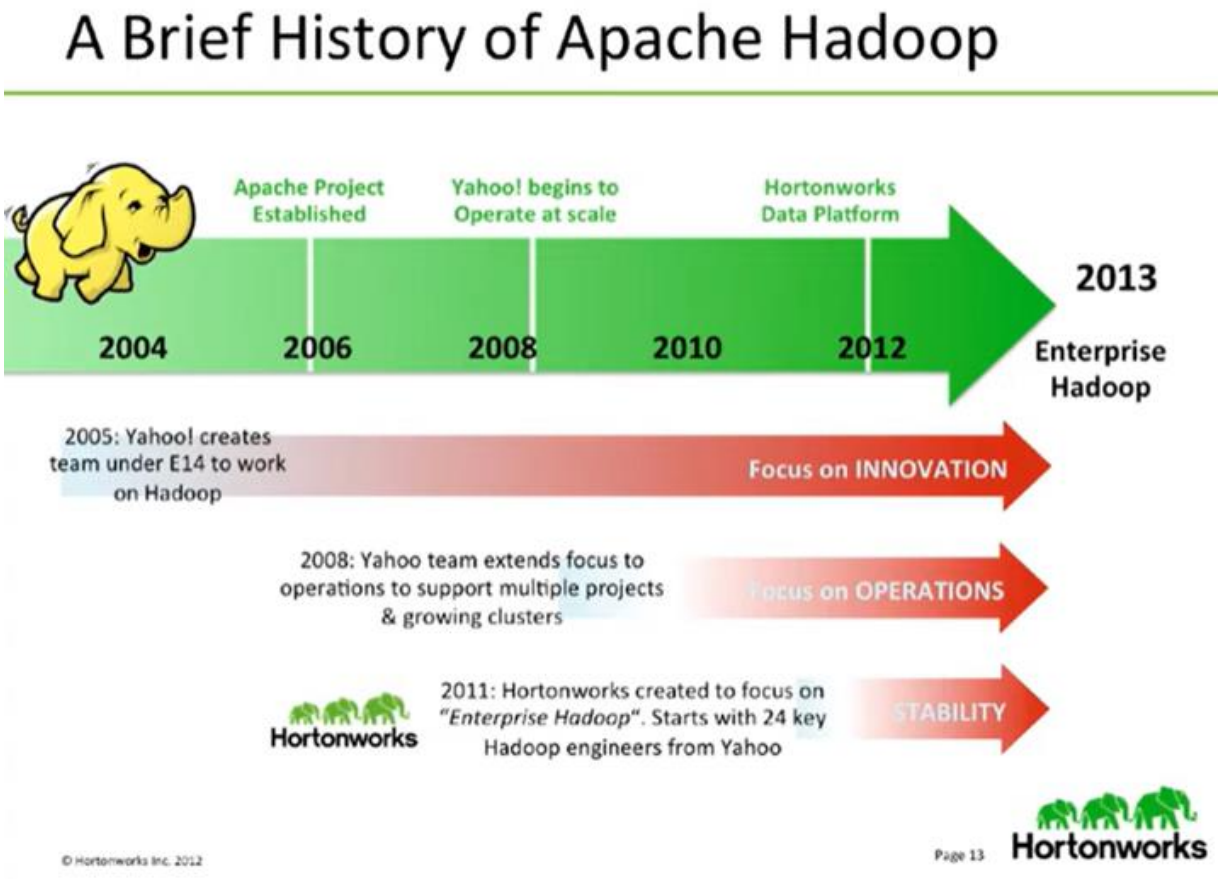
Valorisation de données

Par : Abderrazak Sahraoui

Sommaire Plateformes de Big Data

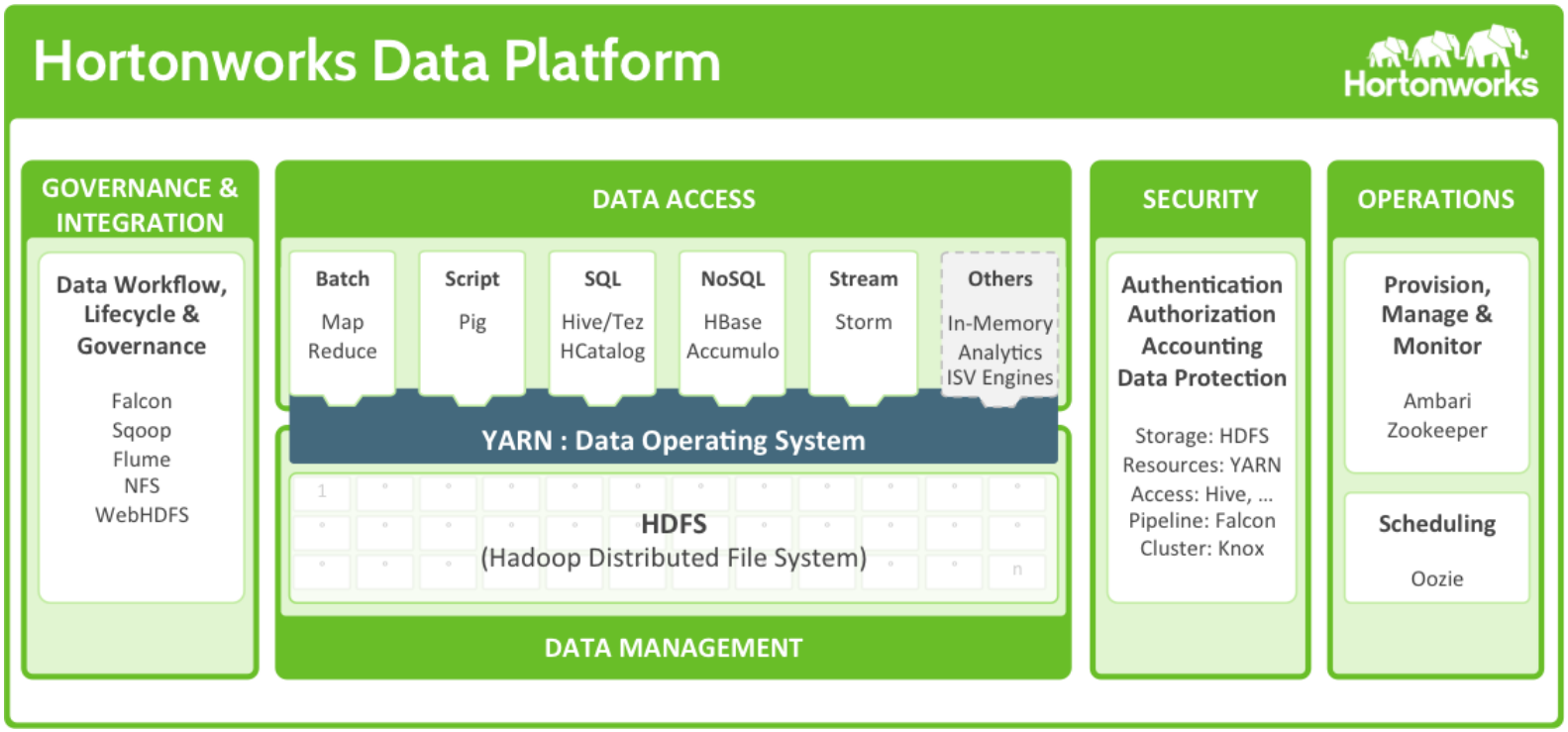
- Plateforme HortonWorks Data Platform (HDP)
- Écosystème Hadoop
- Hadoop vs Bases de données relationnelles
- Architecture de plateforme d'entreprise
- Installation HDP
- Définitions : Ambari, Hcatalog, Hive, Pig, Sparks, Sqoop, Flume, Hbase, Oozie, …
- Autres Plateformes Cloud :
  - Microsoft Azure – HDInsight
  - Amazone - AWS

Hortonworks Data Platform (HDP)

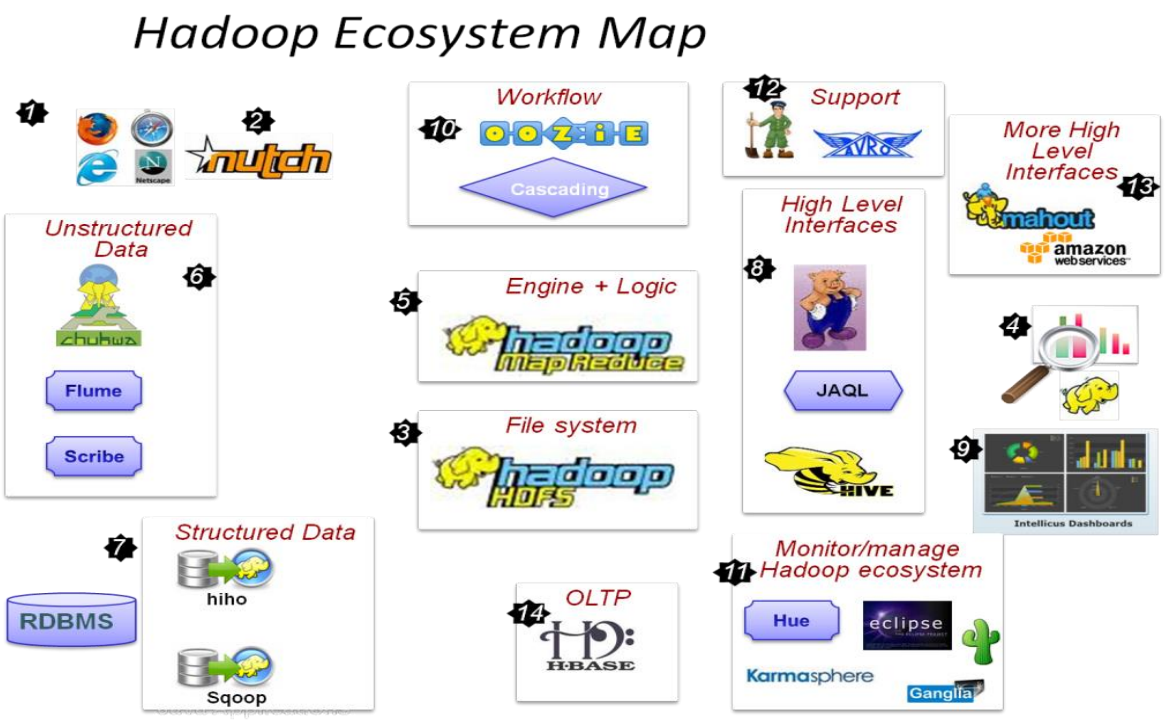


- Lancencement d'une plateforme Hadoop pour entreprise open source en 2012 par Hortonworks

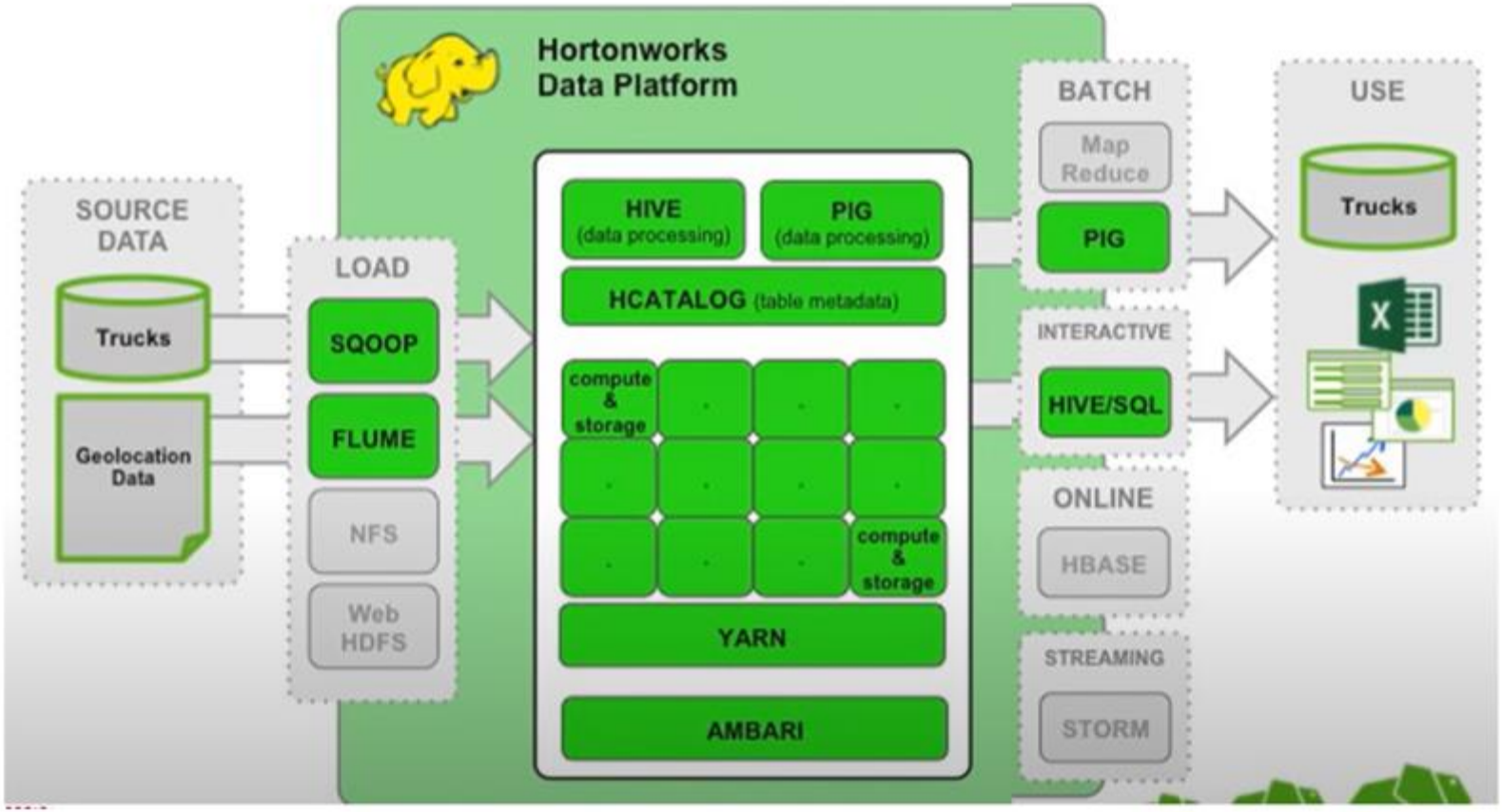
(HDP) Hortonworks Data Platform



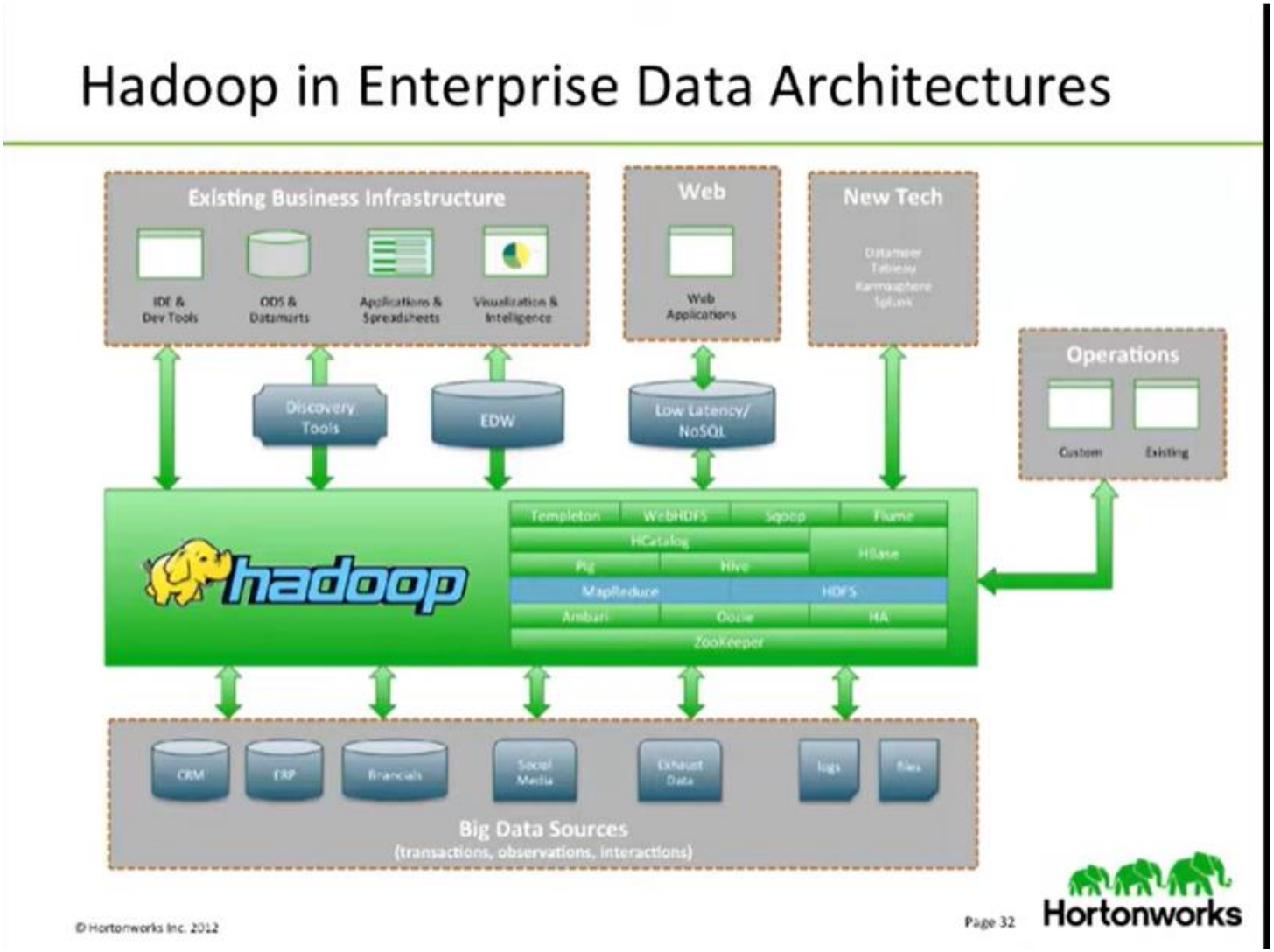
Une autre vision de l'écosystème de Hadoop



Ecosystème d'une plateforme Hadoop



Architecture d'entreprise avec Hadoop



Installation HDP

- Télécharger et installer Oracle Virtual Box
  - https://www.virtualbox.org/
- Exécuter Virtual Box
- Télécharger et installer Hortonworks HDP pour Virtual Box
  - https://www.cloudera.com/downloads/hortonworks-sandbox.html
- Lancer HDP sur Virtual Box
- Ouvrir dans un navigateur web la page : 127.0.0.1:8080

https://www.virtualbox.org/

https://www.cloudera.com/downloads/hortonworks-sandbox.html

Get Started with Hortonworks Sandbox

Hortonworks Sandbox can help you get started learning, developing, testing and trying out new features on HDP and DataFlow.

Hortonworks HDP

The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, Druid and Data Analytics Studio (DAS).

Download Now

Cloudera DataFlow (Ambari)

The Cloudera DataFlow (Ambari)—formerly known as Hortonworks DataFlow—Sandbox makes it easy to get started with Apache NiFi, Apache Kafka, Apache Storm, and Streaming Analytics Manager (SAM).

Download Now

Machine Virtuelle

Oracle VM VirtualBox - Gestionnaire de machines

Fichier Machine Aide

Outils

Nouvelle Ajouter Configuration Oublier Afficher

Hortonworks Sandbox HD...
En fonction

vm
Éteinte

Général

Nom : Hortonworks Sandbox HDP 2.6.5
Système d'exploitation : Red Hat (64-bit)

System

Mémoire vive : 8192 Mo
Processeurs : 4
Ordre d'amorçage : Disque dur, Optique
Accélération : Pagination imbriquée, PAE/NX , Paravirtualisation KVM

Prévisualisation

Affichage

Mémoire vidéo : 8 Mo
Contrôleur graphique : VBoxVGA
Port serveur bureau distant : 5905
Enregistrement : Désactivé

Stockage

Contrôleur : IDE Controller
Maître primaire IDE : Hortonworks Sandbox HDP 2.6.5-disk001.vdi (Normal, 117,19 Gio)

Audio

Pilote hôte : Windows Audio Session
Contrôleur : ICH AC97

Réseau

Interface 1: Intel PRO/1000 MT Desktop (NAT)

USB

127.0.0.1:8080

- Le tableau de bord Ambari s'exécute sur le port :8080 . Par exemple, http://sandbox-hdp.hortonworks.com:8080

Administration HDP par Ambari

Ambari - Sandbox

Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/dashboard/metrics

Ambari Sandbox 0 ops 0 alerts

Dashboard Services Hosts Alerts Admin maria_dev

- HDFS
- YARN
- MapReduce2
- Tez
- Hive
- HBase
- Pig
- Sqoop
- Oozie
- ZooKeeper
- Falcon
- Storm
- Flume
- Ambari Infra
- Atlas
- Kafka
- Knox
- Ranger
- Spark2
- Zeppelin Notebook
- Druid
- Slider

Metrics Heatmaps Config History

Metric Actions Last 1 hour

HDFS Disk Usage	DataNodes Live	HDFS Links	Memory Usage	Network Usage
27%	1/1	NameNode Secondary NameNode 1 DataNodes More...	No Data Available	No Data Available

CPU Usage	Cluster Load	NameNode Heap	NameNode RPC	NameNode CPU WIO
No Data Available	No Data Available	21%	0.08 ms	n/a

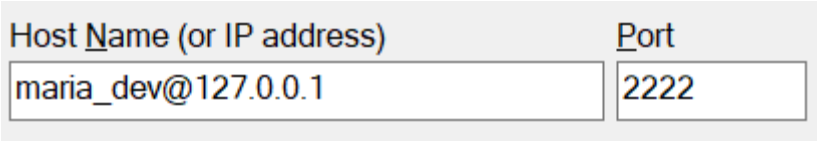
NameNode Uptime	HBase Master Heap	HBase Links	HBase Ave Load	HBase Master Uptime
1.4 hr	n/a	No Active Master 1 RegionServers n/a More...	n/a	n/a

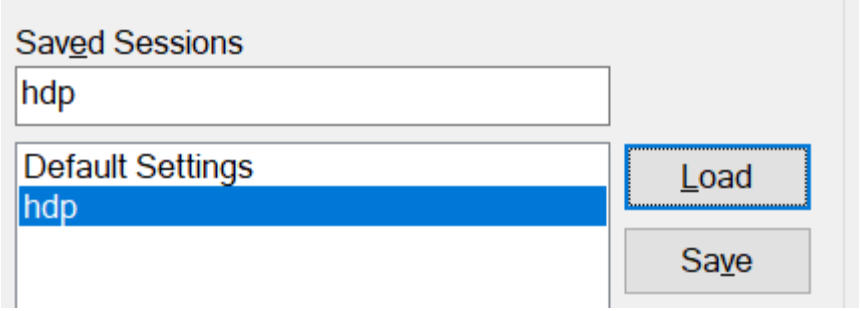
ResourceManager Heap	ResourceManager Uptime	YARN Memory	NodeManagers Live	YARN Links
10%	1.4 hr	0%	1/1	ResourceManager 1 NodeManagers

Tapez ici pour effectuer une recherche FRA 17:14 2023-04-10

Installation Putty

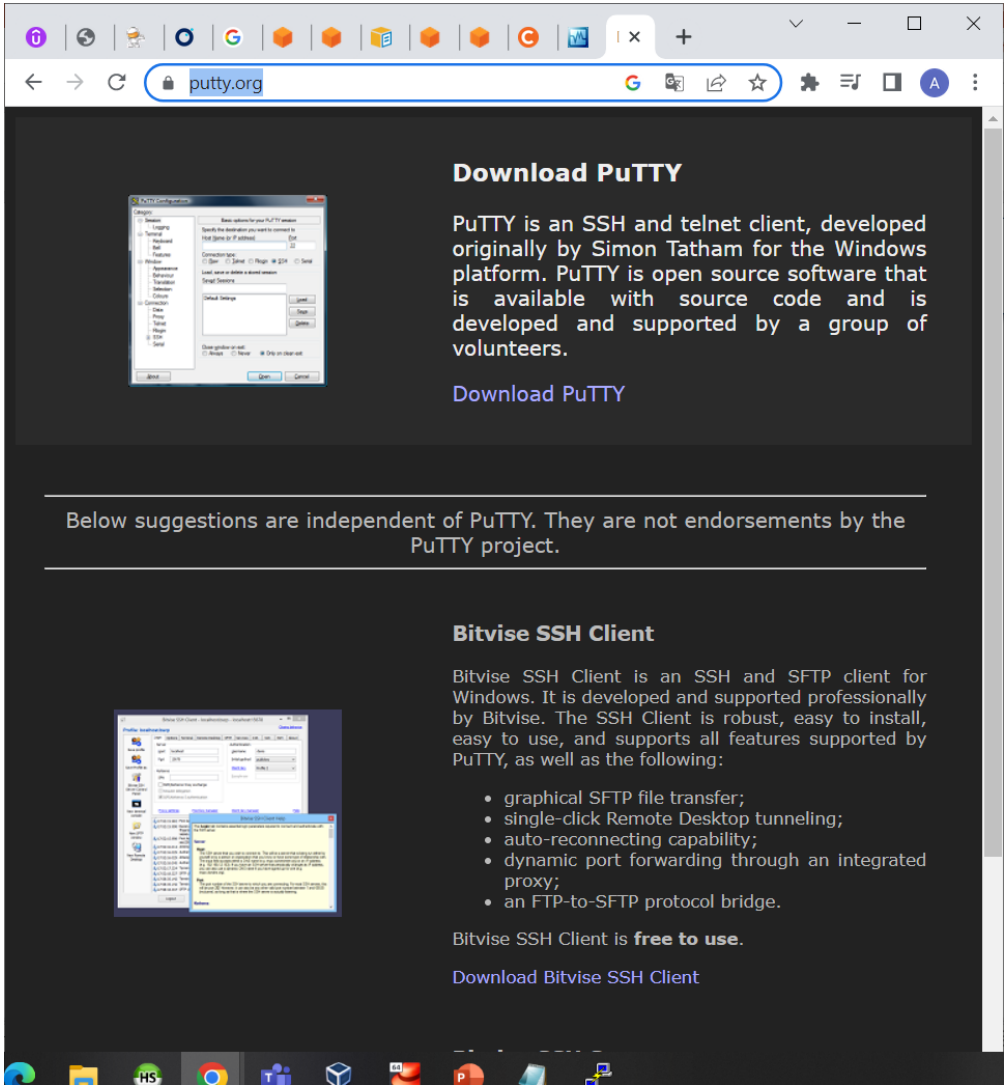
- Télécharger et installer Putty
  - https://www.putty.org/
- Exécuter Putty
- Configurer la ligne HDP



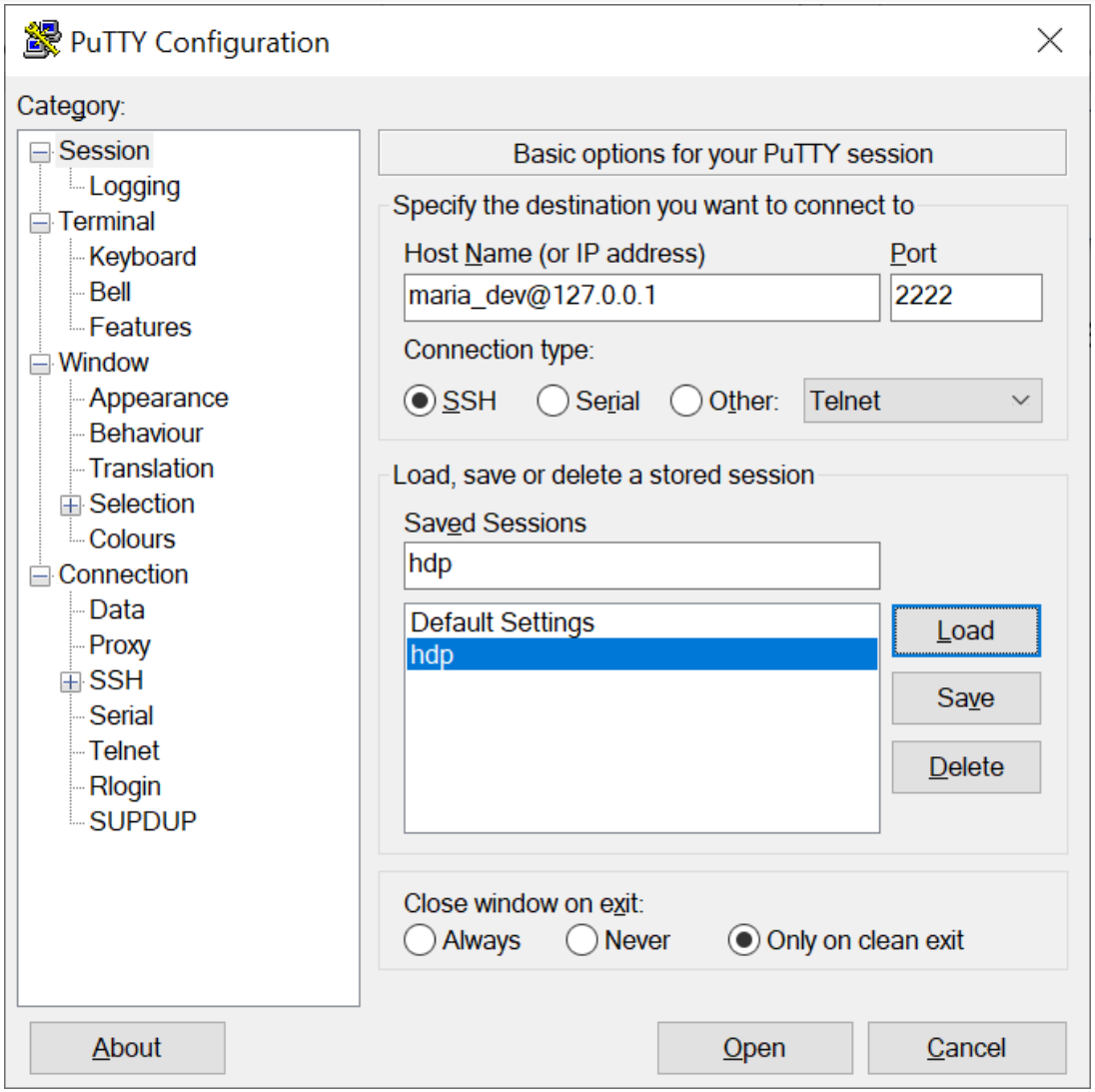


- Cliquer sur Open

https://www.putty.org/



Commande en ligne



Exécution de Putty

maria_dev@sandbox-hdp:~

```
Using username "maria_dev".
maria_dev@127.0.0.1's password:
[maria_dev@sandbox-hdp ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - maria_dev hdfs          0 2023-04-07 21:37 .Trash
drwxr-xr-x   - maria_dev hdfs          0 2023-03-10 00:11 hive
[maria_dev@sandbox-hdp ~]$ ls
[maria_dev@sandbox-hdp ~]$ wget http://media.sundog-soft.com/hadoop/ml-100k/u.data
--2023-04-10 20:35:48--  http://media.sundog-soft.com/hadoop/ml-100k/u.data
Resolving media.sundog-soft.com (media.sundog-soft.com)... 52.217.196.153, 52.217.192.41, 54.
231.228.81, ...
Connecting to media.sundog-soft.com (media.sundog-soft.com)|52.217.196.153|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2079229 (2.0M) [application/octet-stream]
Saving to: 'u.data'

100%[===================================================>] 2,079,229   5.03MB/s   in 0.4s

2023-04-10 20:35:48 (5.03 MB/s) - 'u.data' saved [2079229/2079229]

[maria_dev@sandbox-hdp ~]$
```

Gestion des données sur HDP

- Data Management – Store and process vast quantities of data in a storage layer that scales linearly.

- Hadoop Distributed File System (HDFS) is the core technology for the efficient scale out storage layer, and is designed to run across low-cost commodity hardware

- Apache Hadoop YARN is the pre-requisite for Enterprise Hadoop as it provides the resource management and pluggable architecture for enabling a wide variety of data access methods to operate on data stored in Hadoop with predictable performance and service levels.

Accès aux données sur HDP

- Data Access – Interact with your data in a wide variety of ways – from batch to real-time.
- Apache Hive is the most widely adopted data access technology
- Apache Pig provides scripting capabilities, Apache Storm offers real-time processing, Apache HBase offers columnar NoSQL storage and Apache Accumulo offers cell-level access control.

Accès aux données sur HDP

- Apache Hive – Built on the MapReduce framework, Hive is a data warehouse that enables easy data summarization and ad-hoc queries via an SQL-like interface for large datasets stored in HDFS.

- Apache Pig – A platform for processing and analyzing large data sets. Pig consists of a high-level language (Pig Latin) for expressing data analysis programs paired with the MapReduce framework for processing these programs.

- MapReduce – MapReduce is a framework for writing applications that process large amounts of structured and unstructured data in parallel across a cluster of thousands of machines, in a reliable and fault-tolerant manner.

- Apache Spark – Spark is ideal for in-memory data processing. It allows data scientists to implement fast, iterative algorithms for advanced analytics such as clustering and classification of datasets.

Accès aux données sur HDP

- Apache Storm – Storm is a distributed real-time computation system for processing fast, large streams of data adding reliable real-time data processing capabilities to Apache Hadoop 2.x

- Apache HBase – A column-oriented NoSQL data storage system that provides random real-time read/write access to big data for user applications.

- Apache Tez – Tez generalizes the MapReduce paradigm to a more powerful framework for executing a complex DAG (directed acyclic graph) of tasks for near real-time big data processing.

- Apache Kafka – Kafka is a fast and scalable publish-subscribe messaging system that is often used in place of traditional message brokers because of its higher throughput, replication, and fault tolerance.

- Apache HCatalog – A table and metadata management service that provides a centralized way for data processing systems to understand the structure and location of the data stored within Apache Hadoop.

Accès aux données sur HDP

- Apache Slider – A framework for deployment of long-running data access applications in Hadoop. Slider leverages YARN's resource management capabilities to depl oy those applications, to manage their lifecycles and scale them up or down.

- Apache Solr – Solr is the open source platform for searches of data stored in Hadoop. Solr enables powerful full-text search and near real-time indexing on many of the world's largest Internet sites.

- Apache Mahout – Mahout provides scalable machine learning algorithms for Hadoop which aids with data science for clustering, classification and batch based collaborative filtering.

- Apache Accumulo – Accumulo is a high performance data storage and retrieval system with cell-level access control. It is a scalable implementation of Google's Big Table design that works on top of Apache Hadoop and Apache ZooKeeper.

Gouvernance et Intégration de données

- Data Governance and Integration – Quickly and easily load data, and manage according to policy. Workflow Manager provides workflows for data governance, while Apache Flume and Sqoop enable easy data ingestion, as do the NFS and WebHDFS interfaces to HDFS.

Gouvernance et Intégration de données

- Workflow Management – Workflow Manager allows you to easily create and schedule workflows and monitor workflow jobs. It is based on the Apache Oozie workflow engine that allows users to connect and automate the execution of big data processing tasks into a defined workflow.
- Apache Flume – Flume allows you to efficiently aggregate and move large amounts of log data from many different sources to Hadoop.
- Apache Sqoop – Sqoop is a tool that speeds and eases movement of data in and out of Hadoop. It provides a reliable parallel load for various, popular enterprise data sources.

Gouvernance et Intégration de données

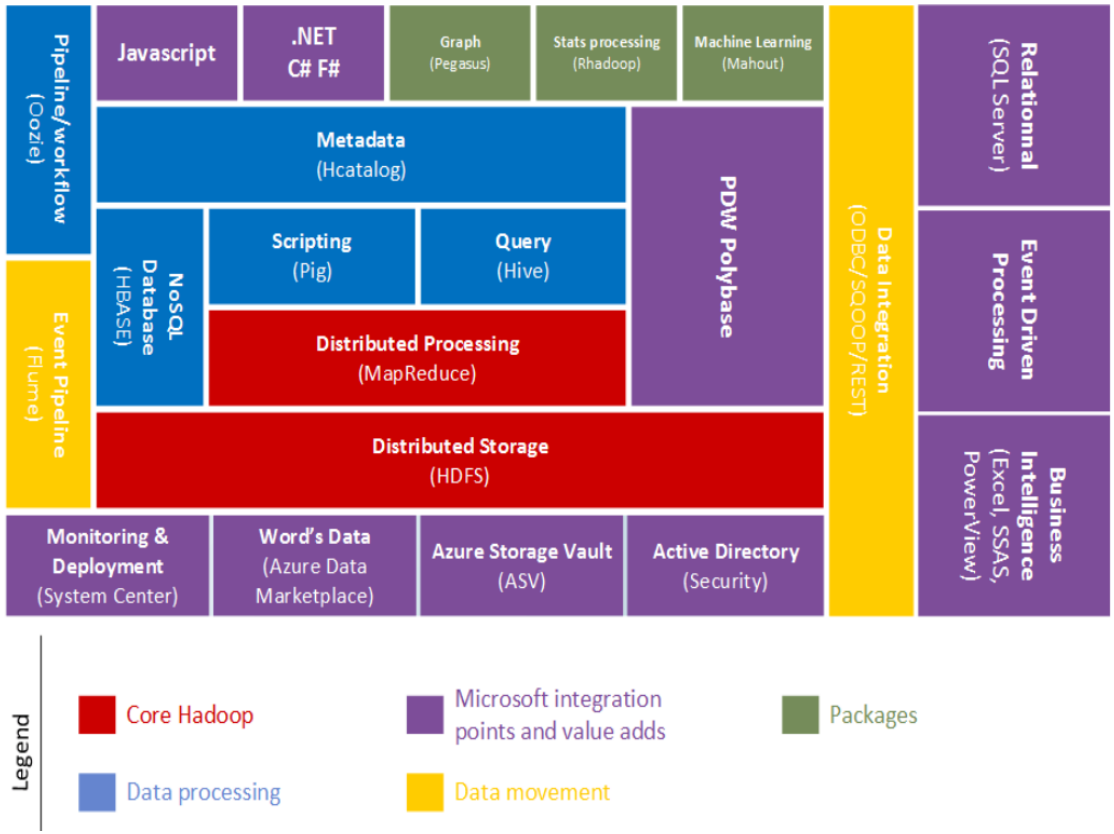
- Security – Address requirements of Authentication, Authorization, Accounting and Data Protection. Security is provided at every layer of the Hadoop stack from HDFS and YARN to Hive and the other Data Access components on up through the entire perimeter of the cluster via Apache Knox.

- Apache Knox – The Knox Gateway ("Knox") provides a single point of authentication and access for Apache Hadoop services in a cluster. The goal of the project is to simplify Hadoop security for users who access the cluster data and execute jobs, and for operators who control access to the cluster.
- Apache Ranger – Apache Ranger delivers a comprehensive approach to security for a Hadoop cluster. It provides central security policy administration across the core enterprise security requirements of authorization, accounting and data protection.

Opérations

- Operations – Provision, manage, monitor and operate Hadoop clusters at scale.
  - Apache Ambari – An open source installation lifecycle management, administration and monitoring system for Apache Hadoop clusters.
  - Apache Oozie – Oozie Java Web application used to schedule Apache Hadoop jobs. Oozie combines multiple jobs sequentially into one logical unit of work.
  - Apache ZooKeeper – A highly available system for coordinating distributed processes. Distributed applications use ZooKeeper to store and mediate updates to important configuration information.

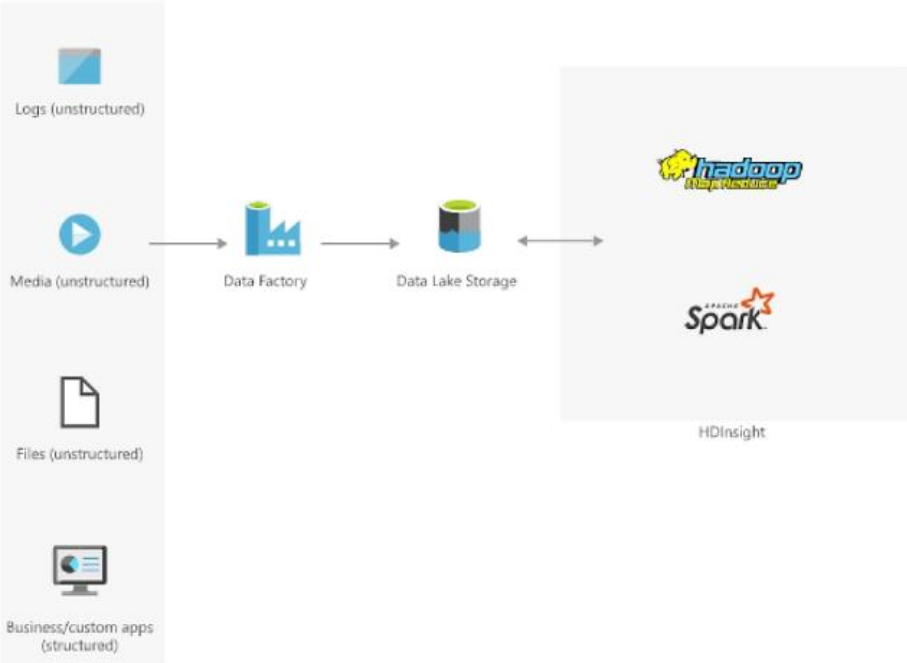
Plateforme Microsoft HDInsight



- HDInsight est basée sur la plateforme Hortonworks
- Les briques violettes dans l'écosystème HDInsight sont les composants ajoutés par Microsoft au produit Hortonworks.

https://blog.octo.com/hdinsight-le-big-data-selon-microsoft/

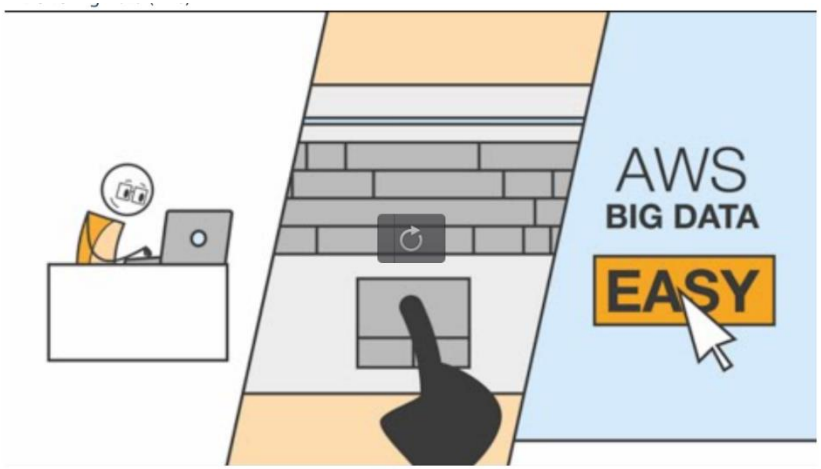
Big data sur Azure



https://azure.microsoft.com/fr-ca/products/hdinsight

- Lac de données grâce à l'intégration des solutions et services de stockage de données Azure.
- Mise à l'échelle automatique en fonction de la charge ou de la planification.
- Tableaux de bord pour surveiller l'intégralité de lac de données à l'aide des

Big data sur Amazon Web Services and Cloud (AWS)



https://aws.amazon.com/fr/big-data/what-is-big-data/

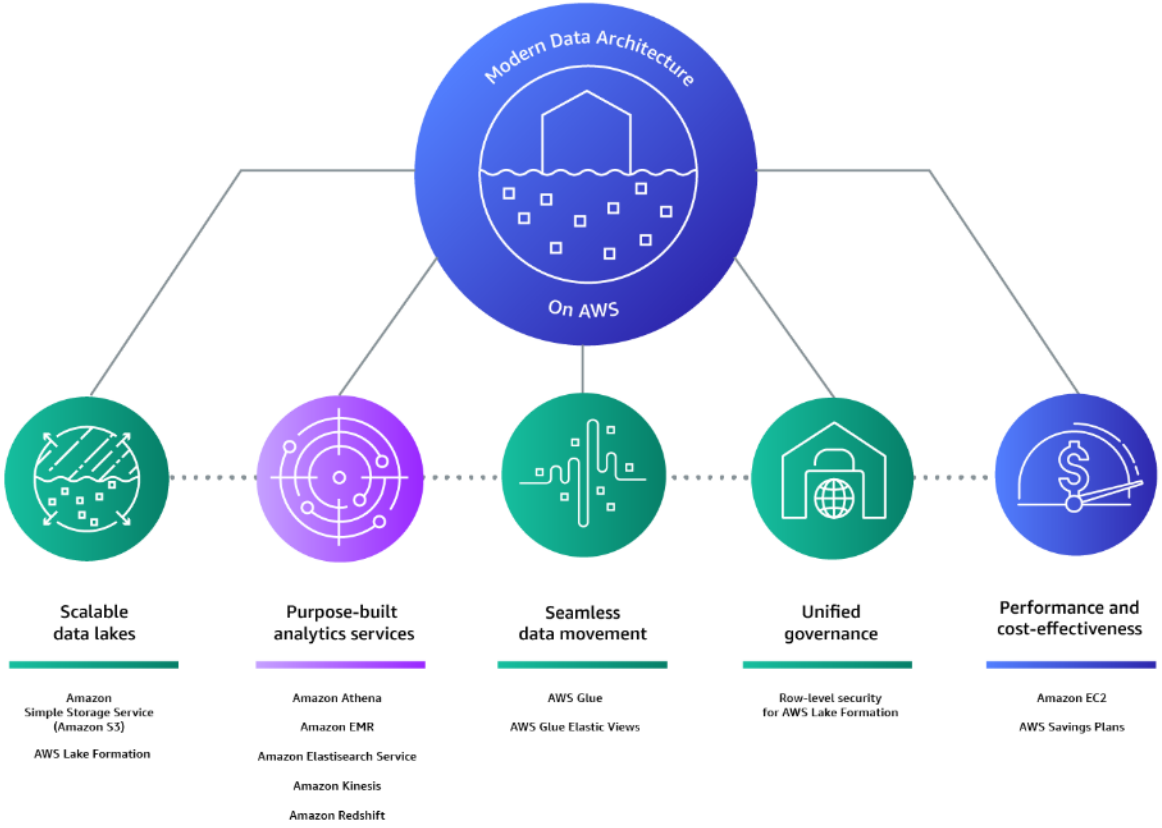
- AWS propose des services dédiés : mouvement des données, stockage des données, lacs de données, analyse des données du big data, analyse des journaux, analyse de streaming, informatique décisionnelle (BI), Apprentissage (machine learning ML).
- Offre clusters et écosystèmes de big data composés de Hadoop, Hive, Pig, Sparks, etc…
- Permet à travers les services EC2 de de monter sa propre solution big data.

Architecture sur AWS

https://aws.amazon.com/fr/big-data/what-is-big-data/

- Les organisations prennent leurs données stockées dans des silos et les déplacent dans un seul endroit pour les analyser et les utiliser pour des tâches de ML (machine learning). Pour réaliser cette opération de façon optimale, elles ont besoin d'utiliser une architecture de données moderne leur permettant de déplacer facilement des données entre des lacs et des magasins de données spécialisés.

Architecture sur AWS



- Intégrer un lac de données, un entrepôt de données et des magasins spécialisés, afin d'unifier la gouvernance et de faciliter le mouvement des données.

https://aws.amazon.com/fr/big-data/what-is-big-data/

Partenaires sur AWS

Cloudera

L'exécution de Cloudera Enterprise sur AWS fournit aux utilisateurs informatiques et professionnels une plateforme de gestion des données qui peut servir de base au traitement et à l'analyse modernes des données.

En savoir plus »

Informatica Cloud

Informatica Cloud offre une intégration optimisée aux services de données AWS avec une connectivité native à plus de 100 applications.

En savoir plus »

Dataguise

Dataguise est le leader de l'exécution métier sécurisée, fournissant des solutions de sécurité centrées sur les données qui détectent et protègent les données sensibles d'une entreprise, peu importe où elles se trouvent ou qui a besoin de les exploiter.

En savoir plus »

Alluxio Data Orchestration

Alluxio Data Orchestration permet aux clients de mieux exploiter les principaux services AWS, tels que EMR et S3 pour les charges de travail d'analytique et d'IA.

En savoir plus »

Amazon EC2

What is Amazon EC2?

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

For more information about cloud computing, see What is cloud computing?

https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/concepts.html

- Monter sa propre solution big data avec EC2