



Une formation
Alphorm.com

Big Data avec Apache Spark

Initiation



Aimad TAHI



Une formation
Alphorm.com

Cursus Big Data avec Spark





Une formation
Alphorm.com

Plan de la formation

Introduction

1. Le Big Data
2. Apache Hadoop
3. Apache Spark
4. Programmer avec Apache Spark
5. Spark SQL
6. Spark Streaming

Conclusion



Une formation
Alphorm.com

Public concerné

Ingénieur d'études et développement
Chef de projet technique
Architecte technique



Prérequis

Connaissances Java, Scala ou Python
Connaissances en SQL
Connaissances d'outils Big Data seront
utiles mais pas obligatoires.

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Présentation du Big Data



Aimad TAHI



Une formation
Alphorm.com

Plan

Le Big Data?

Pourquoi utiliser le Big Data ?

Quelle évolution avec le Big Data ?



Le Big Data c'est quoi ?

Explosion de la quantité des données

Le partage des données

La recherche des données

Le stockage des données

Le traitement des flux de données

Une formation
Alphorm.com



Le Big Data c'est quoi ?

Big Data

- Mégadonnées
- Données massives

Question :

- Big Data = Le volume de données?

Réponse : Faux

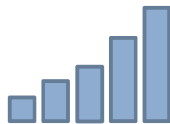
Une formation
Alphorm.com



Le Big Data c'est quoi ?

La règle des 3**V**

Volume



Variété



Vélocité



Une formation
Alphorm.com



Le Big Data c'est quoi ?

Plus 2 **V**

Véracité



Valeur



Une formation
Alphorm.com



Pourquoi ?

Difficulté de traiter la masse de données
produites chaque jour

Les informations sont produites en temps réel

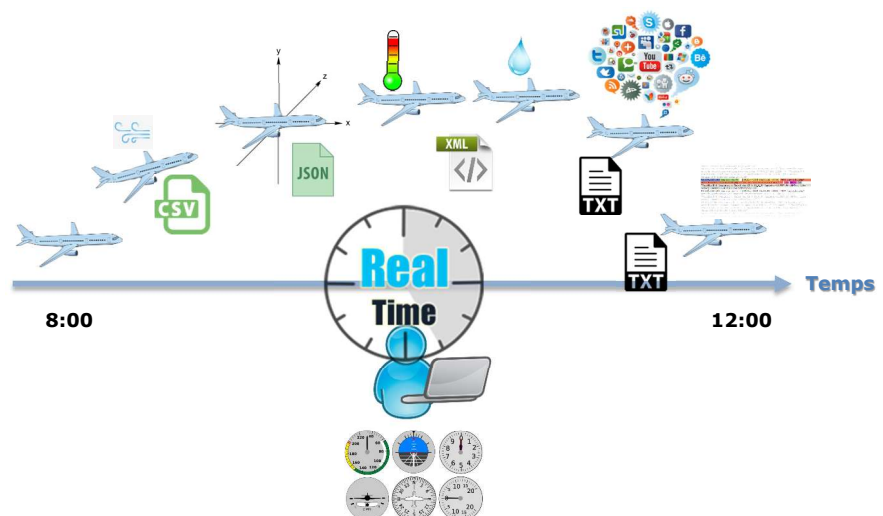
Problème d'optimisation de base de données

Le traitement des données non structurées

Une formation
Alphorm.com



Pourquoi ?



Une formation
Alphorm.com



Pourquoi ?



Une formation
Alphorm.com



Pourquoi ?



Source : http://www.liberation.fr/futurs/2012/12/03/donnees-le-vertige_864585

Une formation
Alphorm.com



Pourquoi ?



Source : http://www.liberation.fr/futurs/2012/12/03/donnees-le-vertige_864585



Evolution avec le Big Data ?

Progression de Hadoop

Une importante réduction des coûts d'analyse des données

Les SGDB NoSQL

La sécurité des données

Le stockage des données sur le cloud

Merci



L'écosystème du Big Data



Aimad TAHI



Systèmes de fichiers



Quantcast File System (QFS)



Une formation
Alphorm.com



Bases NoSQL



elasticsearch



HYPERTABLE INC

Une formation
Alphorm.com



Bases NewSQL



Impala

HCatalog



Une formation
Alphorm.com



Intégration de données



Une formation
Alphorm.com



Programmation distribuée



Une formation
Alphorm.com



Machine Learning



Une formation
Alphorm.com



Visualisation



Une formation
Alphorm.com



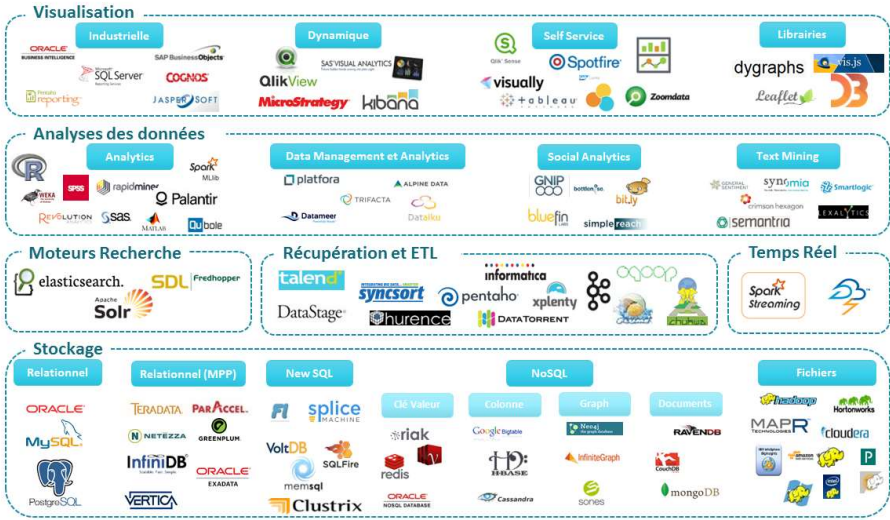
Sécurité



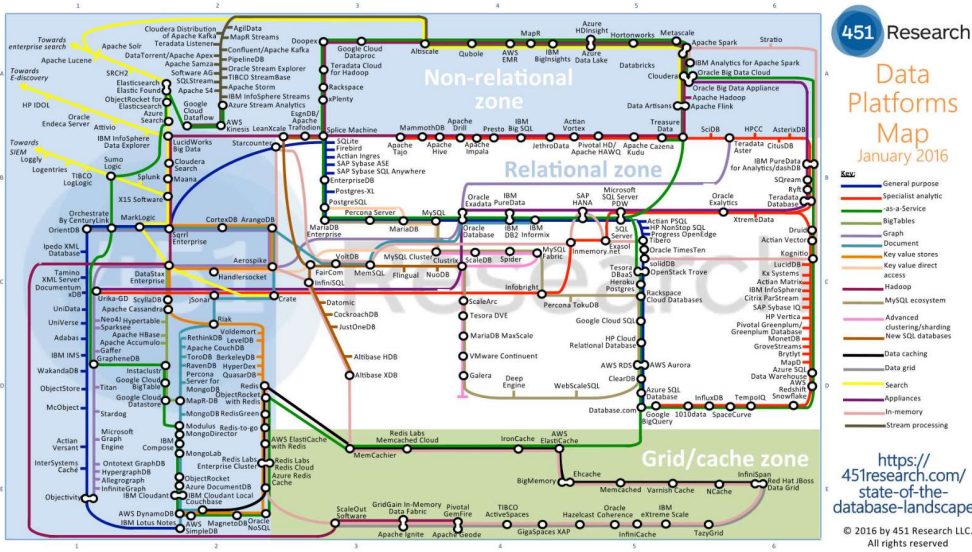
Une formation
Alphorm.com



L'écosystème du Big Data



L'écosystème du Big Data



Tous droits réservés 451 Research

Merci



Présentation Apache Hadoop



Aimad TAHI



Hadoop

Hadoop?

Son utilisation

HDFS (Hadoop Distributed File System)

MapReduce

Distributions Hadoop

Une formation
Alphorm.com



Hadoop?

2004 : conçu par Doug Cutting.

Framework Open Source

Ecrit en java

Aide à créer des applications distribuées

Résiste aux pannes

2009 : fondation logicielle Apache

Une formation
Alphorm.com



Utilisation de Hadoop

Stockage moins coûteux de très gros volumes données (Po)

Les réseaux sociaux   

Les sites e-commerce comme 

Analyse de fichiers non structurés (log)

Une formation
Alphorm.com



HDFS

Hadoop Distributed File System

Système de stockage

Possibilité de stocker des pétaoctets de données

Traitement parallèle et distribué

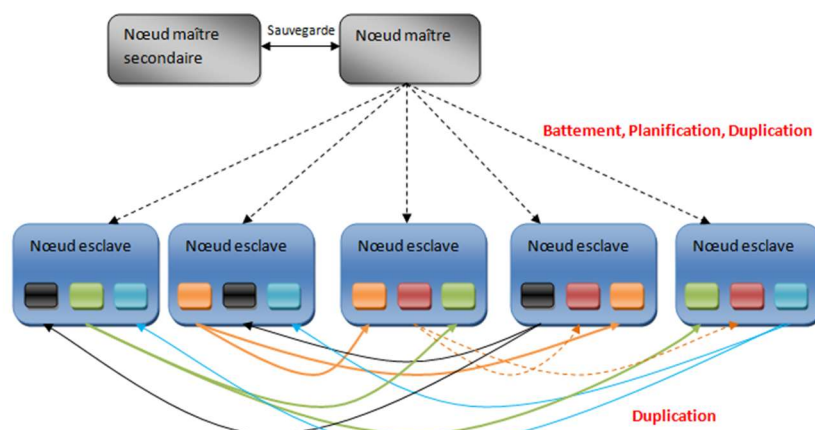
Tolérance aux erreurs avec la réplication des données

Une formation
Alphorm.com



Une formation
Alphorm.com

Schéma de principe du HDFS



<https://fr.wikipedia.org/wiki/Hadoop#/media/File:HDFS.png>



Une formation
Alphorm.com

MapReduce

C'est un modèle de programmation

Calculs parallèles et distribués sur des données très volumineuses

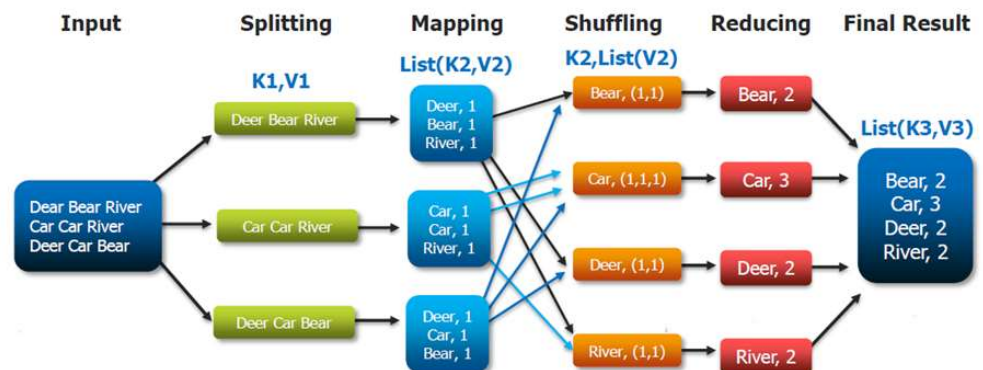
Map réalise des opérations dédiées à chaque élément

Reduce rassemble tous ces éléments et délivre le résultat sous forme condensée



MapReduce

The Overall MapReduce Word Count Process



<https://www.lebigdata.fr/mapreduce-tout-savoir>

Une formation
Alphorm.com



Distributions Hadoop

Versions payantes

Framework Hadoop personnalisé

Consoles de gestion et administration

Documentation, formation ...



Elastic MapReduce

Une formation
Alphorm.com

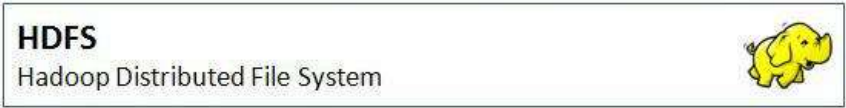
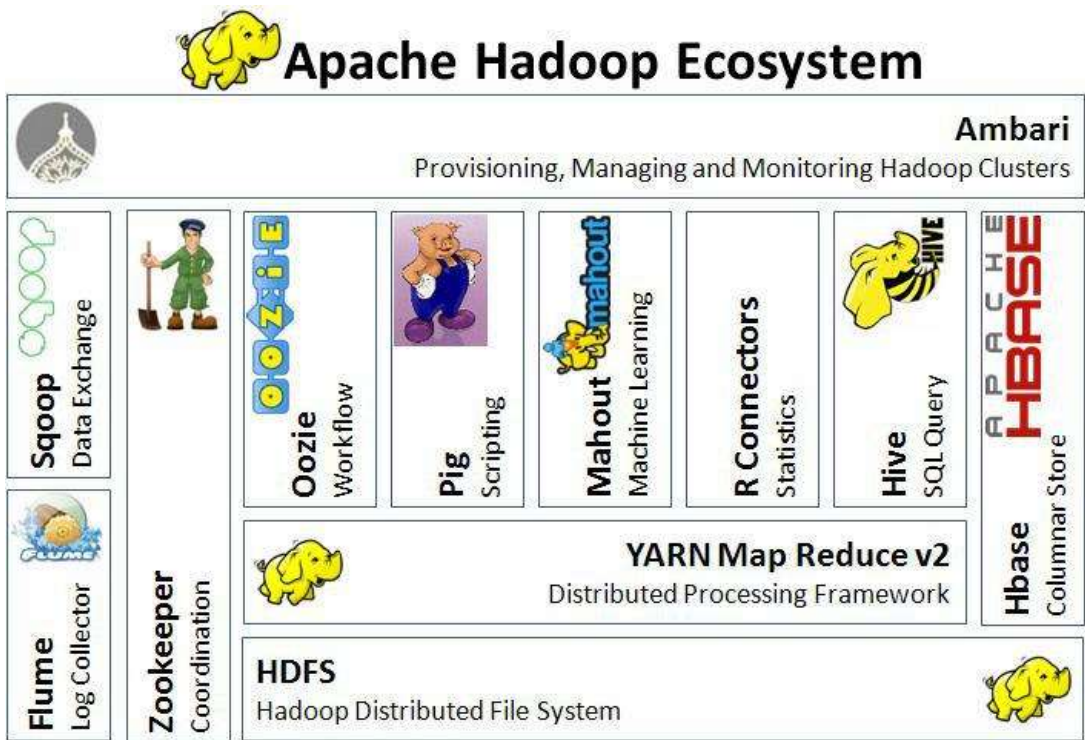
Merci



L'écosystème de Apache Hadoop



Aimad TAHI





Apache Hadoop Ecosystem



YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System



Apache Hadoop Ecosystem



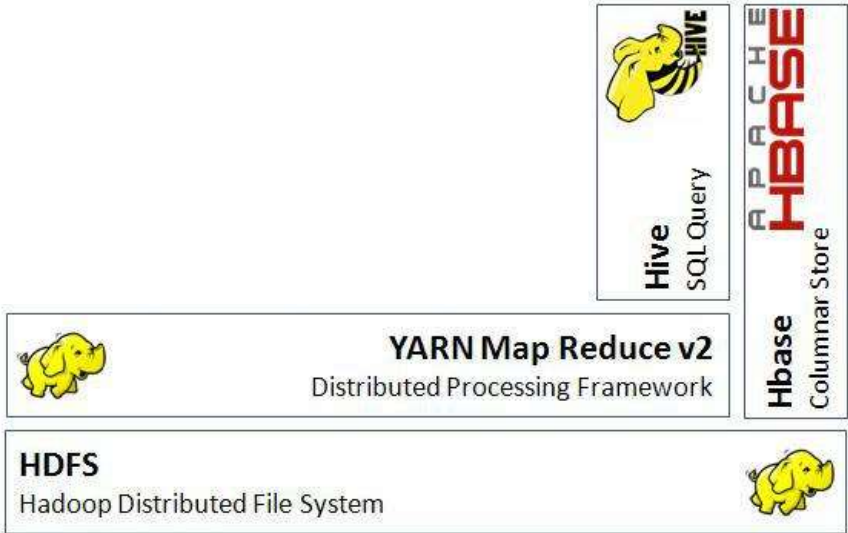
YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

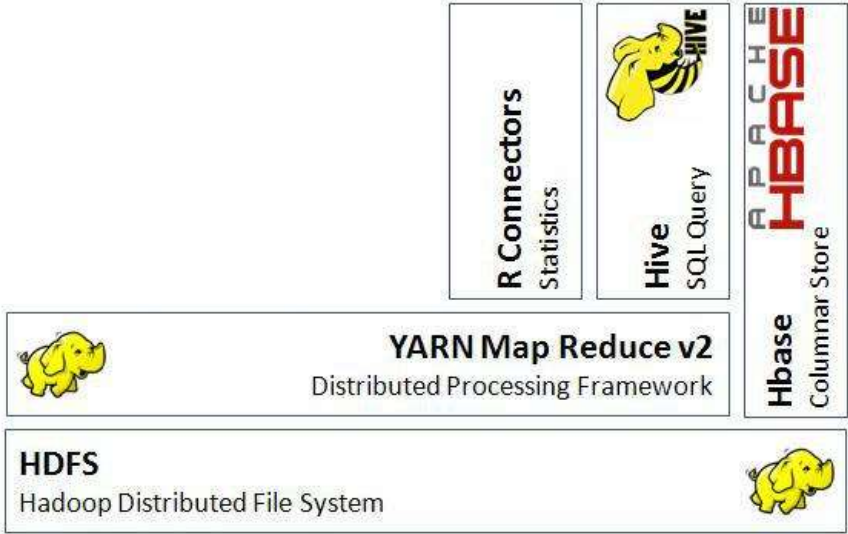


APACHE HBASE
Hbase
Columnar Store

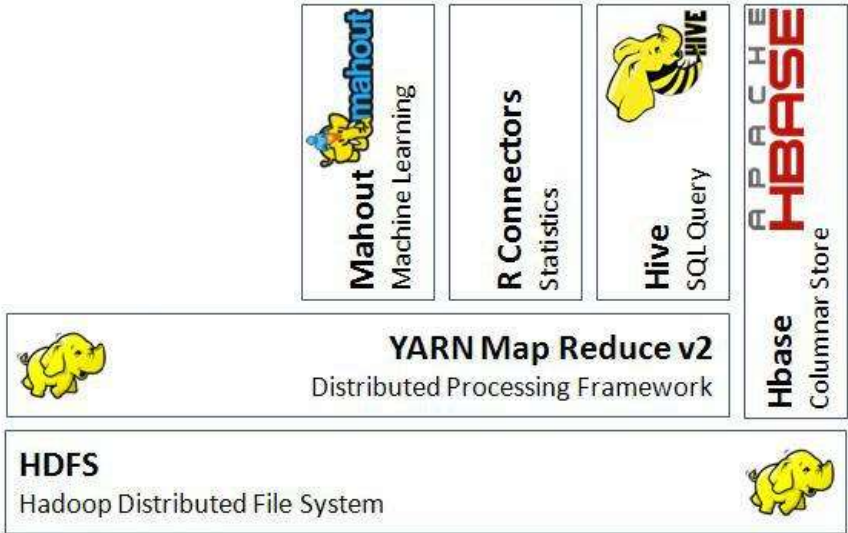
 Apache Hadoop Ecosystem



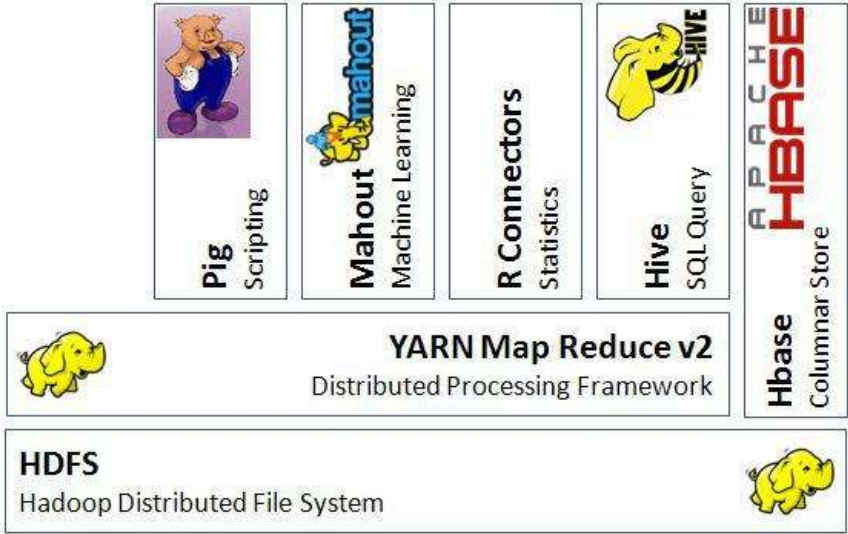
 Apache Hadoop Ecosystem



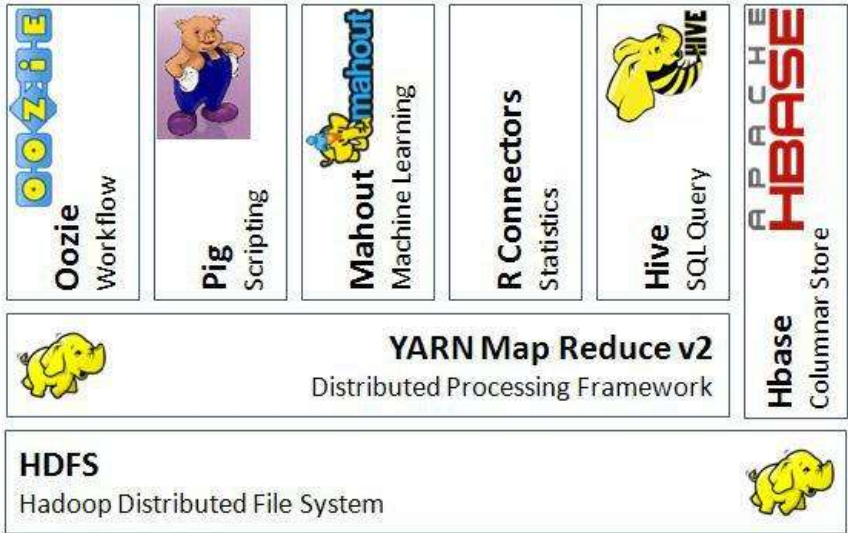
 Apache Hadoop Ecosystem



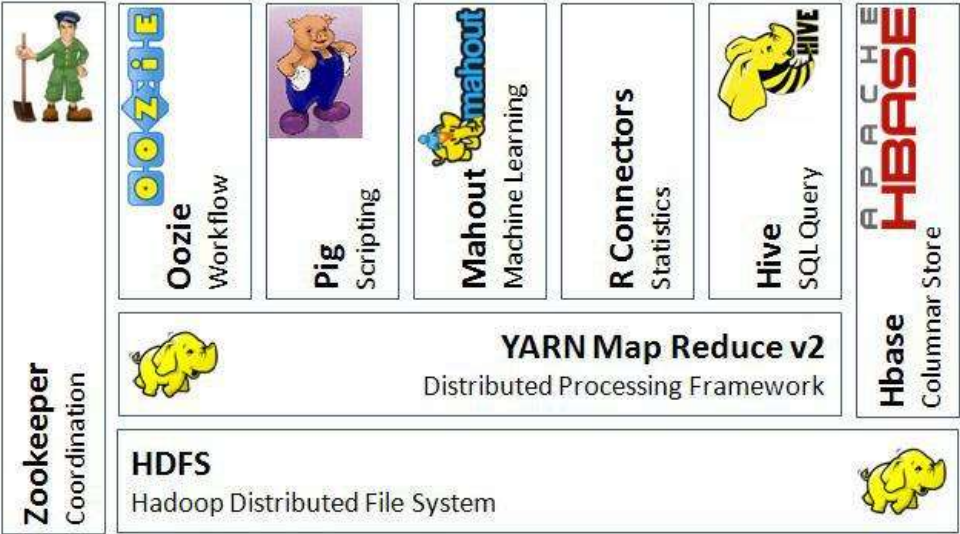
 Apache Hadoop Ecosystem



 Apache Hadoop Ecosystem

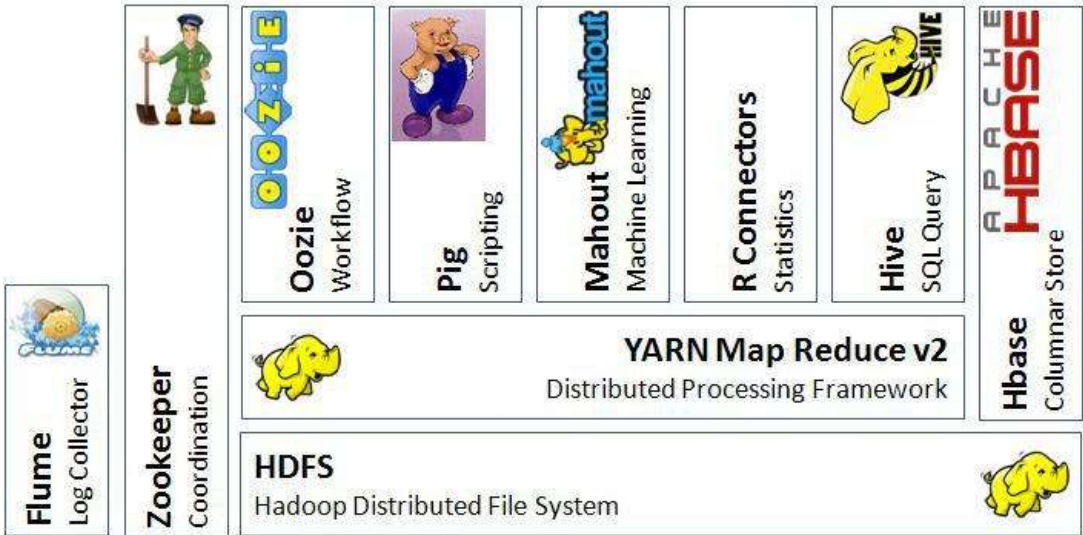


 Apache Hadoop Ecosystem

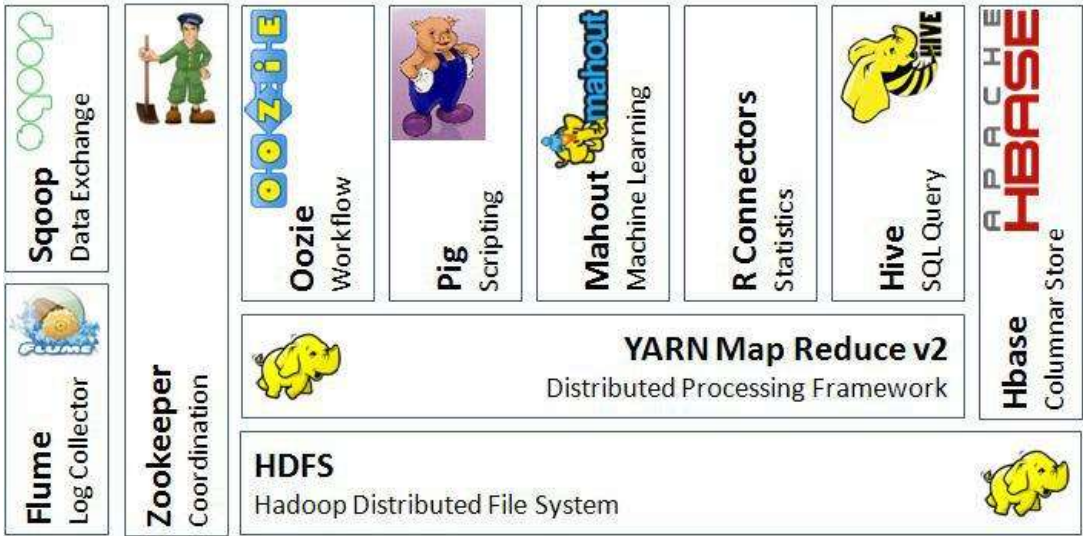


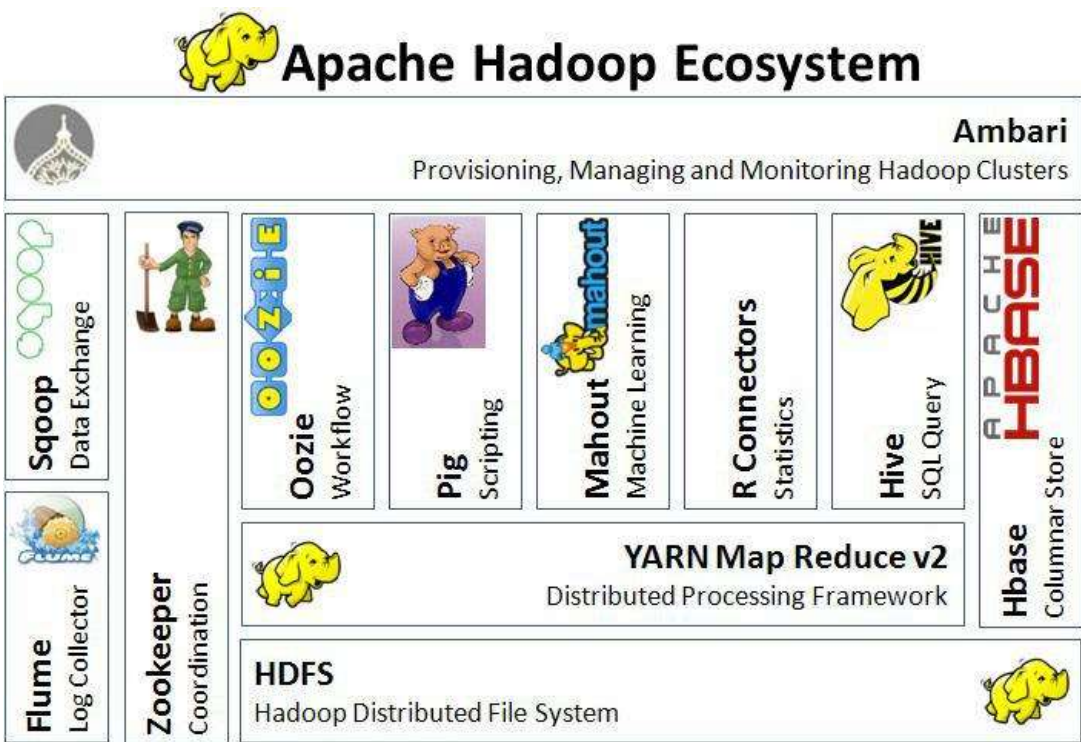


Apache Hadoop Ecosystem



Apache Hadoop Ecosystem





Merci



Une formation
Alphorm.com

Présentation d'Apache Spark



Aimad TAHI



Une formation
Alphorm.com

Plan

Spark?

Pourquoi utiliser Spark ?

Applications utilisant Spark?

Interfaces de programmation Spark



Spark?

2009 : Création par Matei Zaharia

Laboratoire AMPLab Université de Berkeley

2010 : Open Source sous licence BSD

2013 : Licence Apache Software Foundation

2014 : ASF Top Level Project

L'entreprise Databricks

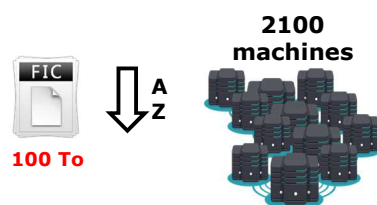


Une formation
Alphorm.com

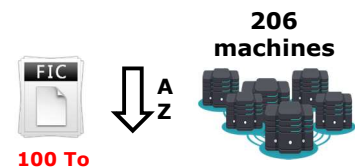


Spark?

2014 : Spark a gagné le Daytona GraySort Contest



72 Minutes



23 Minutes

Une formation
Alphorm.com



Pourquoi utiliser Spark ?

3 fois plus rapide que Hadoop
Utilise moins de ressources que Hadoop
Plus de 400 contributeurs comme
*Intel, Facebook, IBM, Netflix, Yahoo!,
Databricks ...*

Une formation
Alphorm.com



Pourquoi utiliser Spark ?

Fonctionne en mémoire (in-memory)
Possibilité d'utiliser Spark avec Hadoop
Résilient aux pannes

Une formation
Alphorm.com



Une formation
Alphorm.com

Applications utilisant Spark?

Les ETL

Analyse prédictive et Machine Learning

Opérations d'accès aux données (SQL)

Traitement et extraction de texte

Traitement temps réel

Applications graphiques



Une formation
Alphorm.com

API Spark

Il est écrit en Scala

Fonctionne sous JVM

Fournit un support natif pour les APIs :

- Scala
- Python
- Java
- SQL
- R

Merci



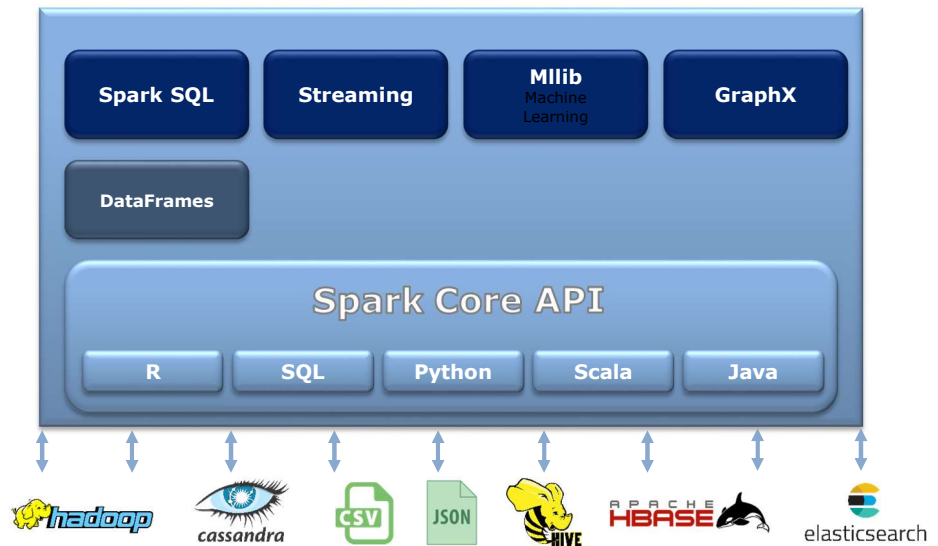
L'écosystème D'Apache Spark



Aimad TAHI



Écosystème Spark



Une formation
Alphorm
com

Merci



Une formation
Alphorm.com

Installation de Spark sous Red Hat/CentOS



Aimad TAHI



Une formation
Alphorm.com

Installation de Spark sous Microsoft Windows



Aimad TAHI



Une formation
Alphorm.com

Installation de Spark sous Mac OS X



Aimad TAHI



Une formation
Alphorm.com

Installation de Spark sous Ubuntu/Debian



Aimad TAHI



Installation

ubuntu[®]



Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Comprendre l'architecture des applications de Spark



Aimad TAHI



Une formation
Alphorm.com

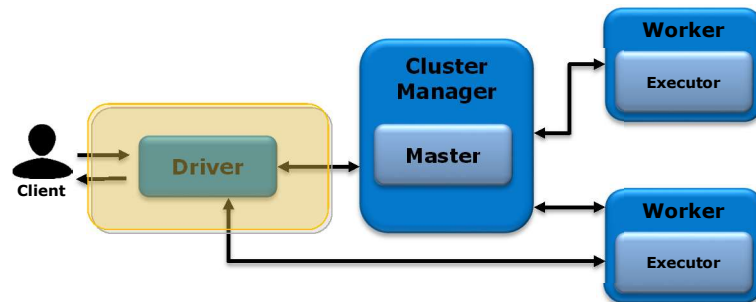
Plan

Architecture Master/Worker
Spark Driver
Spark Executors et Workers
Spark Master et Cluster Manager (CM)



Une formation
Alphorm.com

Architecture Master/Worker



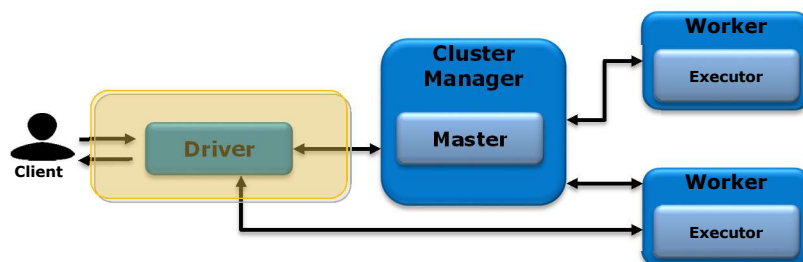
Une formation
Alphorm.com

Spark Driver

Spark Context

Planification d'application

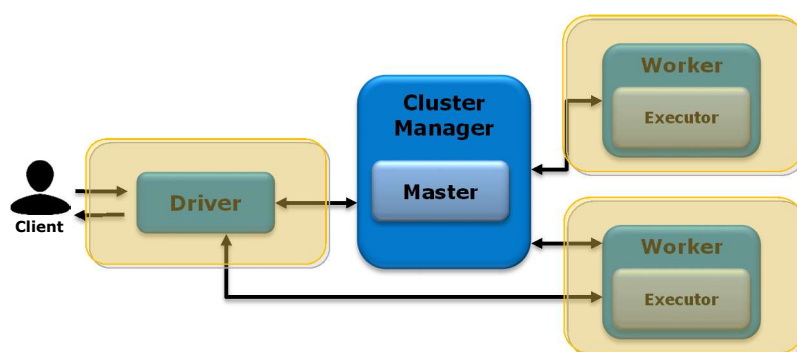
Ordonnanceur d'application





Une formation
Alphorm.com

Architecture Master/Worker



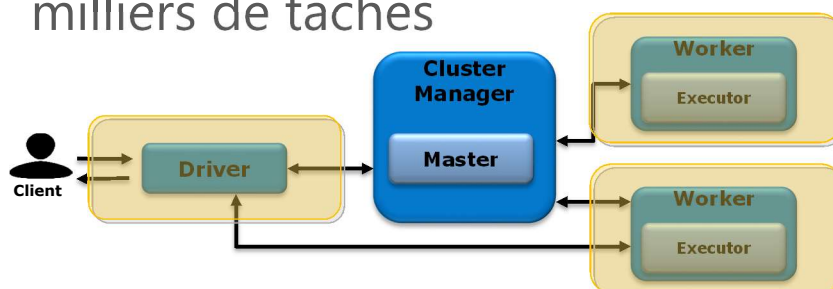
Une formation
Alphorm.com

Spark Executors et Workers

Réservation CPU

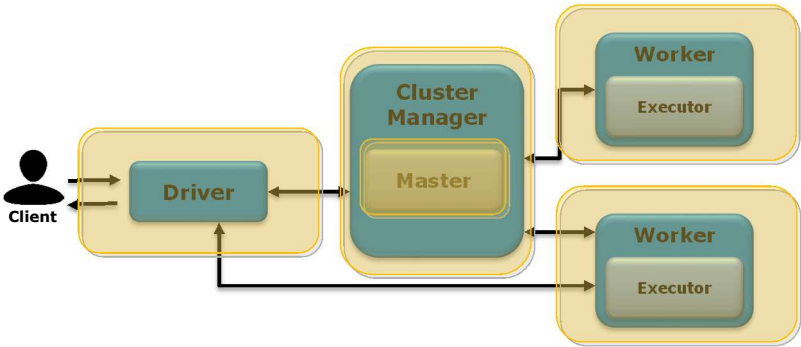
Réservation mémoire

Un Exécuteur peut exécuter des
milliers de tâches

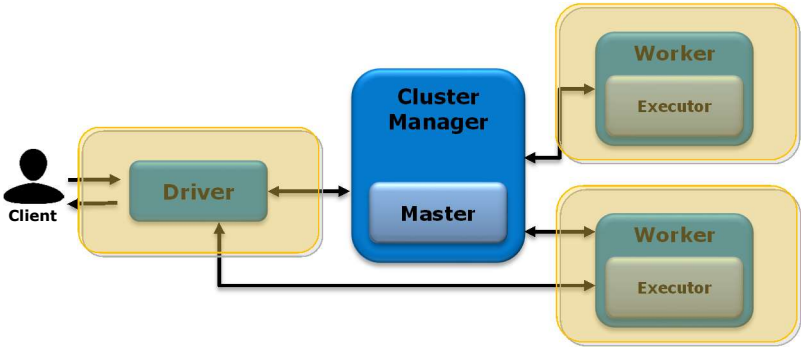




Architecture Master/Worker



Architecture Master/Worker



Merci



Programmation fonctionnelle dans Spark



Aimad TAHI



Une formation
Alphorm.com

Plan

Introduction
Les fonctions
Exemple



Une formation
Alphorm.com

Introduction

Peu de lignes de code
Enchaîner les fonctions dans la même ligne de code
Codes plus lisibles
Réutilisabilité



Les fonctions

Affectation d'une fonction à une variable

Une fonction peut être passée en paramètre d'une autre

La récursivité (fonction factorielle)



Exemple

```
def opp(x : Int) : Int = { return -x; }
```

```
def opp(x : Int) : Int = -x
```



Exemple

```
def max(x : Int, y : Int) = if (x > y) x else y
```



Exemple

Code scala

```
class Person( val prenom: String, val nom: String){  
  def sepresenter : String = "Bonjour, je suis "  
    + prenom + " "  
    + nom  
}
```



Exemple

Code java

```
public class Person {  
    private final String prenom;  
    private final String nom;  
    public Person(String prenom, String nom){  
        this.prenom = prenom;  
        this.nom = nom;  
    }  
    public String getprenom(){  
        return prenom;  
    }  
    public String getnom(){  
        return nom;  
    }  
    public String sepresenter () {  
        return (Bonjour, je suis " + prenom + " " + nom);  
    }  
}
```

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Scala avec Spark



Aimad TAHI



Une formation
Alphorm.com

Plan

Introduction

REPL

Les variables et valeurs

Case class

Immutabilité



Introduction

Hybride Objet / fonctionnel

Compilé en bytecode

S'exécute dans la JVM

Scalable

Concis (Programmation fonctionnelle et typage)

Une formation
Alphorm.com



REPL

Read-Eval-Print loop

Ligne de commande permettant de faire du « pas à pas »

Parfait pour l'apprentissage et la mise au point

Une formation
Alphorm.com



Une formation
Alphorm.com

Les variables et valeurs

Une variable déclarée avec "val" ne peut pas être ré-affectée

Une variable déclarée avec "var" peut être ré-affectée

On parle de la variable : pas de ses éventuels attributs



Une formation
Alphorm.com

Case class

Des classes dont les objets sont immuables

Pas de new

Comparaison par valeur



Immutabilié

Objet dont l'état ne peut pas être modifié après sa création

Optimisation au niveau du compilateur

Merci



Une formation
Alphorm.com

Scala

Types, Liste et Tableaux



Aimad TAHI



Une formation
Alphorm.com

Plan

Types
Liste
Tableaux



Une formation
Alphorm.com

Types

Int
Double
String
Char
Byte
BigInt
...



Une formation
Alphorm.com

Liste

Les listes sont immutables (= on ne peut pas changer le contenu)

List[String] est une liste de chaîne de caractères



Tableaux

Les éléments d'un tableaux sont mutables

Initialisation

Modification

Affichage

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Scala

Fonctions, Currying et Tuple



Aimad TAHI



Une formation
Alphorm.com

Plan

Fonctions
Currying
Tuple



Fonctions

Mot-clé def (Déclaration d'une fonction)

Le nom de la fonction

Les paramètres de la fonction

Le type de retour

Le corps de la fonction entre accolades

Une formation
Alphorm.com



Currying

Fonction où seuls quelques arguments sont spécifiés

Concepts courants dans les langage fonctionnels

Une formation
Alphorm.com



Tuple

Nombre fixe d'éléments pouvant être
manipulés d'un seul bloc

Merci



Une formation
Alphorm.com

Scala

Pattern matching et les Classes



Aimad TAHI



Une formation
Alphorm.com

Plan

Pattern matching
Classes



Pattern matching

Comme le switch
Puissant

Une formation
Alphorm.com



Classes

Les paramètres deviennent des membres publics à la classe

scala.Seq est le super type avec les méthodes :

- filter
- fold
- map
- reduce
- ...

Une formation
Alphorm.com

Merci



Scala Les collections



Aimad TAHI



Plan

Collections

Manipulation des collections

Une formation
Alphorm.com



Les collections

Une syntaxe compacte pour les
Tableaux, List, Set, Map, etc

Inférence de type

Versions mutables et immutables

Une formation
Alphorm.com



Manipulation des collections

Un grand choix de fonctions
Consistant sur l'ensemble des collections
foreach, map, filter, partition, find
fold, reduce, etc...

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Introduction aux RDD



Aimad TAHI



Une formation
Alphorm.com

Plan

RDD?
Resilient
Distributed
Dataset



RDD?

C'est une structure de données centrale de Spark

C'est du code fourni par le framework Spark

Ce code appartient à la librairie du framework dans chacun des langages suivants :

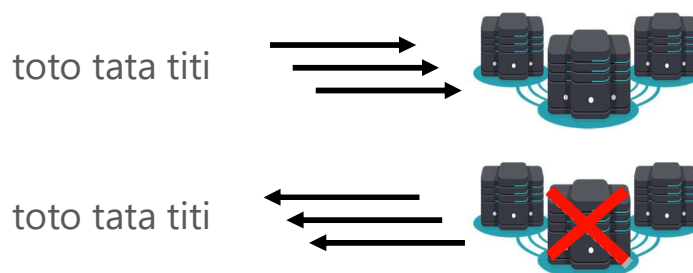


Une formation
Alphorm.com



Resilient

Tolérance aux pannes matérielles.



Une formation
Alphorm.com



Distributed

Opérations effectuées d'une manière distribuée

Données distribuées sur les nœuds du cluster

Une formation
Alphorm.com



Dataset

Opérations sur une abstraction de collection

Une formation
Alphorm.com

Merci



Apprendre les bases de la programmation des RDD



Aimad TAHI



Une formation
Alphorm.com

Plan

Création d'un RDD

Structure d'un RDD

Distribution et partitions

Les opérations sur un RDD

Transformation et DAG

Action



Une formation
Alphorm.com

Création d'un RDD

SparkContext (sc) permet de créer les RDD

- A partir d'une source de données
- En parallélisant une collection
- En appliquant une opération de transformation



Création d'un RDD

Méthode **textFile("/chemin_du_fichier")**

```
//Local
```

```
val rdd1 = sc.textFile("/user/myfiles")
```

```
//Hadoop
```

```
val rdd2 = sc.textFile("hdfs://user/myfiles")
```

Une formation
Alphorm.com



Création d'un RDD

Méthode **parallelize(liste)**

```
//création d'une liste en scala
```

```
val liste = List("bleu", "rouge", "vert")
```

```
val rdd3 = sc.parallelize(liste)
```

Une formation
Alphorm.com



Une formation
Alphorm.com

Structure d'un RDD

Immutable : Ne peut pas être modifiée

Lazy : Contient des références aux données

Cacheable : Un RDD peut être sauvegarder



Une formation
Alphorm.com

Distribution et partitions

Lors de sa création, un RDD est partitionné

Le nombre de partitions est dépendant de
plusieurs paramètres

Les partitions sont parfois imposées

Par défaut



Les opérations sur un RDD

Les Transformations

- sont lazy

Les Actions

- Un RDD va être évalué une fois l'action est exécutée

Une formation
Alphorm.com

 = easy

 = medium

Essential Core & Intermediate Spark Operations

TRANSFORMATIONS

General

- map
- filter
- flatMap
- mapPartitions
- mapPartitionsWithIndex
- groupBy
- sortBy

Math / Statistical

- sample
- randomSplit

Set Theory / Relational

- union
- intersection
- subtract
- distinct
- cartesian
- zip

Data Structure / I/O

- keyBy
- zipWithIndex
- zipWithUniqueId
- zipPartitions
- coalesce
- repartition
- repartitionAndSortWithinPartitions
- pipe

ACTIONS

- reduce
- collect
- aggregate
- fold
- first
- take
- forEach
- top
- treeAggregate
- treeReduce
- forEachPartition
- collectAsMap

- count
- takeSample
- max
- min
- sum
- histogram
- mean
- variance
- stdev
- sampleVariance
- countApprox
- countApproxDistinct

- takeOrdered

- saveAsTextFile
- saveAsSequenceFile
- saveAsObjectFile
- saveAsHadoopDataset
- saveAsHadoopFile
- saveAsNewAPIHadoopDataset
- saveAsNewAPIHadoopFile

 = easy

 = medium

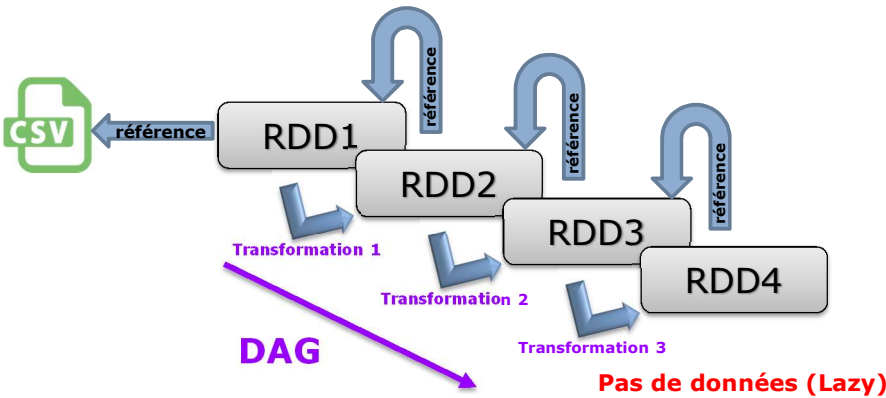
Essential Core & Intermediate PairRDD Operations

	General	Math / Statistical	Set Theory / Relational	Data Structure
	• flatMapValues • groupByKey • reduceByKey • reduceByKeyLocally • foldByKey • aggregateByKey • sortByKey • combineByKey	• sampleByKey	• cogroup (=groupWith) • join • subtractByKey • fullOuterJoin • leftOuterJoin • rightOuterJoin	• partitionBy
		• keys • values	• countByKey • countByValue • countByValueApprox • countApproxDistinctByKey • countApproxDistinctByKey • countByKeyApprox • sampleByKeyExact	

<https://training.databricks.com/visualapi.pdf>



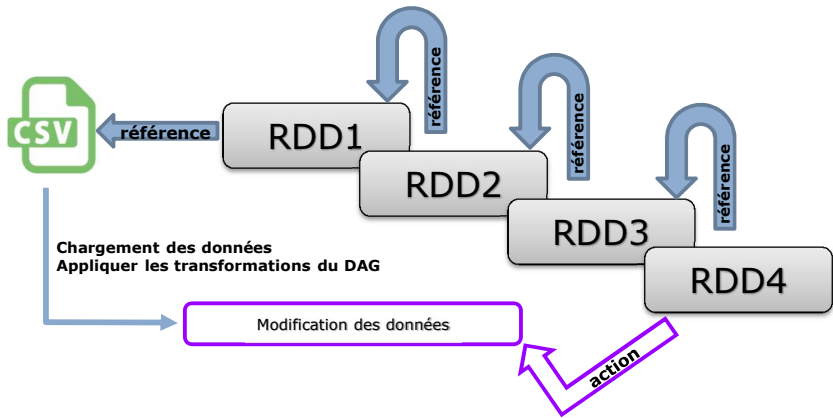
Transformation et DAG



*DAG = Directed Acyclic Graph



Action



Merci



Une formation
Alphorm.com

Comprendre les concepts de MapReduce



Aimad TAHI



Une formation
Alphorm.com

Plan

Inconvénients Hadoop MapReduce
MapReduce avec Spark
Pourquoi en RAM ?
MapReduce versus Spark
Exemple Wordcount en Spark



Hadoop MapReduce

Inconvénients ☹️

Résultat écrit sur le disque

Opérations Map Reduce se font sur les disques

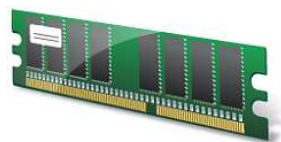


Une formation
Alphorm.com



MapReduce avec Spark

Écriture des résultats en RAM

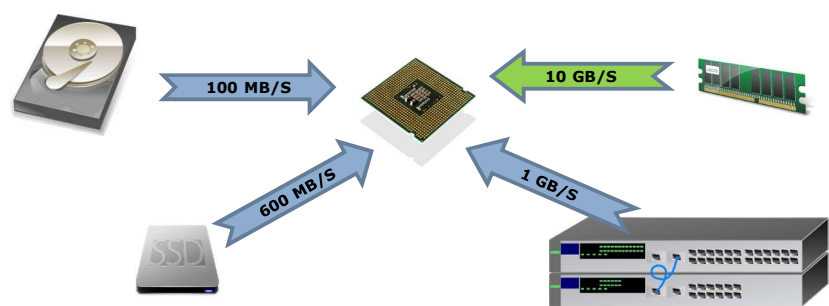


Une formation
Alphorm.com



Pourquoi en RAM ?

Taux de transfert de données avec le processeur

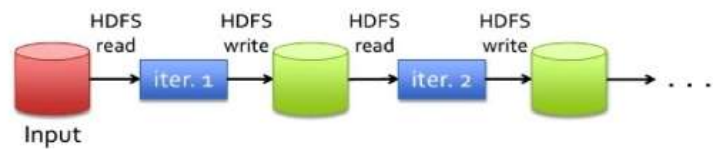


Une formation
Alphorm.com



MapReduce versus Spark

MapReduce (Hadoop)



Spark



Une formation
Alphorm.com



Wordcount en Spark



Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Les opérations générales



Aimad TAHI



Une formation
Alphorm.com

Plan

Les Transformations :

- map, filter , flatmap,
- groupBy, sortBy,
- groupByKey, reduceByKey, sortByKey

Les Actions :

- reduce,
- fold, collect,
- keys, values



Les transformations

Méthode **map**

Méthode **filter**

Méthode **flatMap**

Une formation
Alphorm.com



Les transformations

Méthode **groupBy**

Méthode **sortBy**

Une formation
Alphorm.com



Les transformations

Méthode **groupByKey**

Méthode **reduceByKey**

Méthode **sortByKey**

Une formation
Alphorm.com



Les Actions

Méthode **reduce**

Méthode **fold**

Méthode **collect**

Une formation
Alphorm.com



Les Actions

Méthode **keys**

Méthode **values**

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Les opérations Math/Statistical



Aimad TAHI



Une formation
Alphorm.com

Plan

Les Transformations :

- sample

Les Actions :

- count



Les transformations

Méthode **Sample**



Une formation
Alphorm.com



Les actions

Méthode **count**



Une formation
Alphorm.com

Merci



Les opérations relationnelles



Aimad TAHI



Une formation
Alphorm.com

Plan

Les Transformations :

- union, intersection, subtract
- distinct, cartesian, cogroup
- join, rightOuterJoin, leftOuterJoin

Les Actions :

- takeOrdered



Une formation
Alphorm.com

Les transformations

Méthode **union**

Méthode **intersection**

Méthode **subtract**



Les transformations

Méthode **distinct**

Méthode **cartesian**

Méthode **cogroup**

Une formation
Alphorm.com



Les transformations

Méthode **join**

Méthode **rightOuterJoin**

Méthode **leftOuterJoin**

Une formation
Alphorm.com



Les Actions

Méthode **takeOrdered**

Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Les opérations Data Structure



Aimad TAHI



Une formation
Alphorm.com

Plan

Les Transformations :

- keyBy, coalesce

Les Actions :

- saveAsTextFile



Les transformations

Méthode **keyBy**

Méthode **coalesce**

Une formation
Alphorm.com



Les Actions

Méthode **saveAsTextFile**

Une formation
Alphorm.com

Merci



Introduction à Spark SQL



Aimad TAHI



Une formation
Alphorm.com

Plan

Introduction

Les composants de Spark SQL

- DataFrame
- SQLContext, HiveContext



Une formation
Alphorm.com

Introduction

Composant du framework Apache Spark

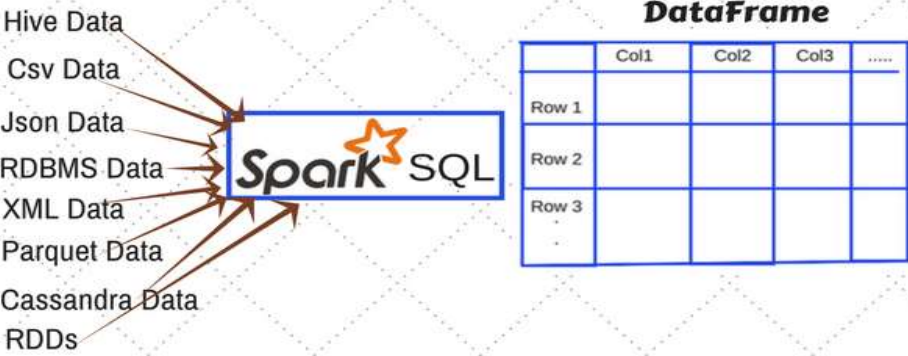
Interface de programmation permettant de requêter
des sources de données ayant un schéma

Permet de traiter des données stockées sous
différents formats (Json, sgbd ...)



Introduction

Ways to Create DataFrame in Spark



DataFrame

Collection de données distribuées





DataFrame

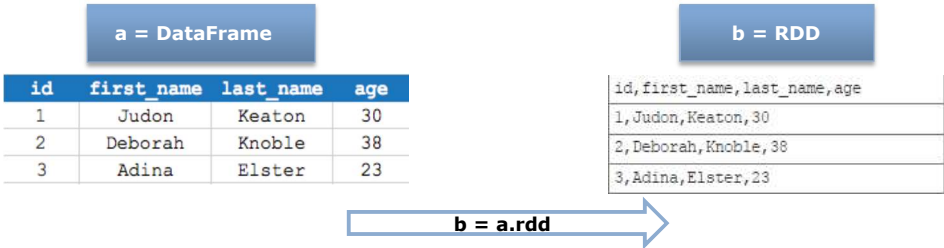
Données organisées en colonnes nommées

id	first_name	last_name	age
1	Judon	Keaton	30
2	Deborah	Knoble	38
3	Adina	Elster	23
4	Antoinette	Campe	75
5	Cob	Larmouth	52
6	Isador	Westwood	49



DataFrame

Peuvent être converties en RDD





DataFrame

Peuvent être créées à partir des :

- RDD
- fichiers de données structurées
- données JSON
- tables Hive
- Bases de données externes

Une formation
Alphorm.com



SQLContext

Créés à partir du SparkContext

```
val sqlContext = new org.apache.spark.sql.SQLContext(sc)
```

Point d'entrée pour manipuler des données via
SparkSQL

Gère le cycle de vie des dataframes

Une formation
Alphorm.com



HiveContext

Créés à partir du SparkContext

```
import org.apache.spark.sql.hive.HiveContext  
val hiveContext = new HiveContext(sc)
```

Manipuler des données via SQL

Supporte tous les formats de stockage de Hive
(Parquet, ORC, Avro, RCFile)

La connexion à hive est paramétrée dans

```
$SPARK_HOME/CONF/hive-site.xml
```

Merci



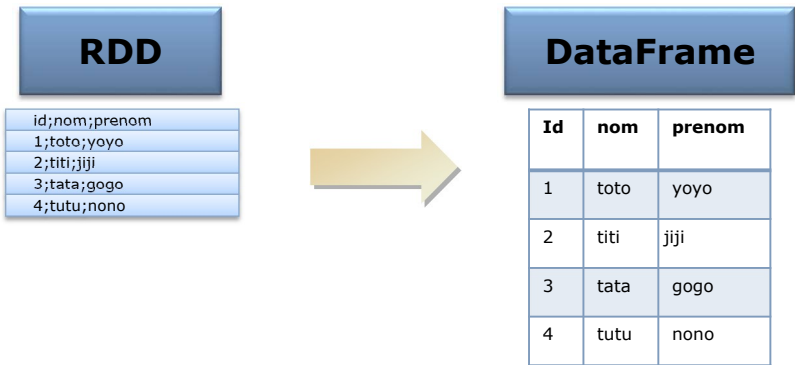
Création d'un DataFrame à partir d'un RDD



Aimad TAHI



DataFrame à partir d'un RDD



Merci



Création d'un DataFrame à partir d'un Objet JSON



Aimad TAHI



DataFrame à partir d'un objet JSON



```
{"name":"Michael"}  
{"name":"Andy", "age":30}  
{"name":"Justin", "age":19}
```

DataFrame

name	age
Michael	
Andy	30
Justin	19

Merci



Une formation
Alphorm.com

Création de DataFrames à partir de fichiers



Aimad TAHI



Une formation
Alphorm.com

DataFrames à partir de fichiers



DataFrame		
c1	c2	c3
1	toto	yoyo
2	titi	jiji
3	tata	gogo
4	tutu	nono

Merci



Convertir un DataFrame en RDD



Aimad TAHI



Convertir un DataFrame en RDD

DataFrame

Id	nom	prenom
1	toto	yoyo
2	titi	jiji
3	tata	gogo
4	tutu	nono



RDD

id;nom;prenom
1;toto;yoyo
2;titi;jiji
3;tata;gogo
4;tutu;nono

Merci



Une formation
Alphorm.com

Utiliser Spark SQL avec les SGBDR



Aimad TAHI



Une formation
Alphorm.com

Bases de données relationnelles

Une base de données relationnelle peut-être considérée comme une source de données pour Spark SQL :

- Pour lire les données
- Aussi pour les écrire



Exemple JDBC



Une formation
Alphorm.com

Merci



Une formation
Alphorm.com

Utilisation des DataFrame avec Spark SQL



Aimad TAHI



Une formation
Alphorm.com

Plan

Opérations sur les métadonnées
Opérations de base



Opérations sur les métadonnées

DataFrame.columns

DataFrame.dtypes

Une formation
Alphorm.com



Opérations de base

show()

drop()

filter()

distinct()

Une formation
Alphorm.com

Merci



Utilisation des Dataset avec Spark SQL



Aimad TAHI



Une formation
Alphorm.com

Plan

Limitation des RDD

Limitation des Dataframe

Dataset

Performance

Caching memory



Une formation
Alphorm.com

Limitation des RDD

Pas de schéma de données

Pas de moteur d'optimisation



Limitation des Dataframe

- Pas de sécurité de type à la compilation
- Pas de conversion objet de domaine



Dataset

RDD	Dataframe	Dataset
données non structuré	Traitement structuré	Structuré et non structuré
API Fonctionnel	API Relationnelle	API fonctionnelle et relationnelle
Version 0.1	Version 1.3	Version 1.6



Dataset

Collection distribuée de données

Offre les avantages des RDD

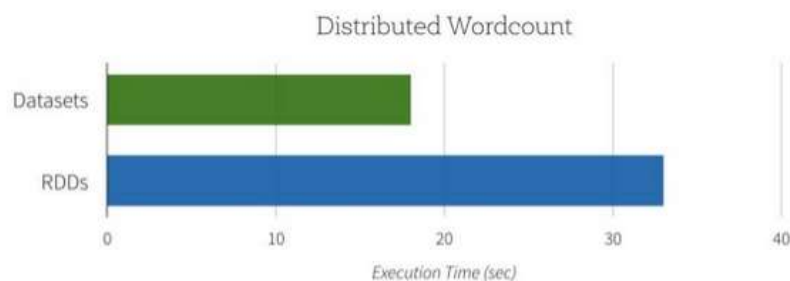
Offre les avantages du moteur d'exécution optimisé de Spark SQL.

Utilise un encodeur spécialisé pour sérialiser les objets.

Une formation
Alphorm.com



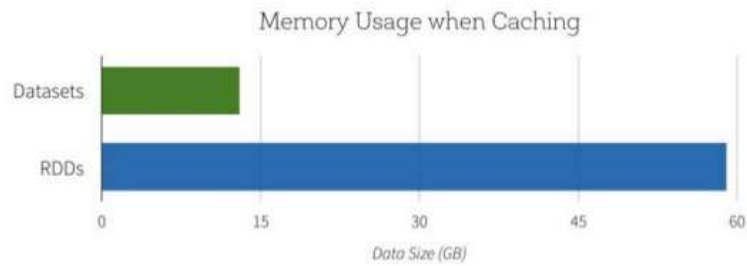
Performance



Une formation
Alphorm.com



Caching memory



Merci



Une formation
Alphorm.com

Présentation de Spark Streaming



Aimad TAHI



Une formation
Alphorm.com

Plan

Introduction

Architecture

DStream

Les Transformations DStream

Window et sliding



Une formation
Alphorm.com

Introduction

Extension de Spark, permettant la gestion de gros volumes de données arrivant en continu

Tolérant aux pannes

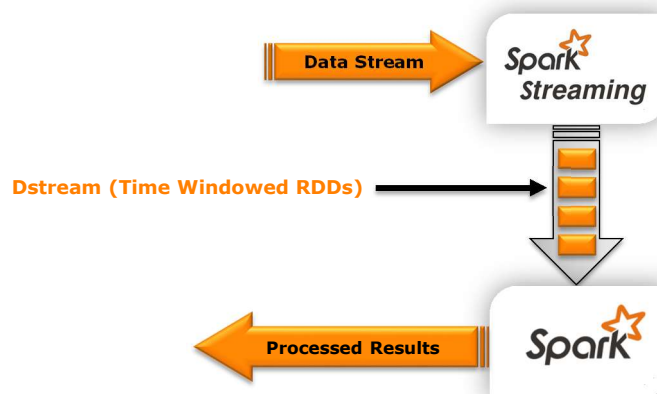


Une formation
Alphorm.com

Architecture

En entrée : Data Stream

En sortie : Processed Results





DStream

Discretized Streams (DStreams)

Représente un flux continu de données



Une formation
Alphorm.com



DStream

Séquence de RDD

- ✓ Partitionnées
- ✓ Immutables
- ✓ Tolérance aux pannes



Une formation
Alphorm.com



Une formation
Alphorm.com

Les Transformations DStream

Plusieurs transformations disponibles via l'API RDD

Les transformations DStream créent un nouveau DStream (comme les RDD)

Transformation :

- ✓ map(func)
- ✓ flatMap(func)
- ✓ filter(func)
- ✓ reduce(func) ...



Une formation
Alphorm.com

Window et sliding

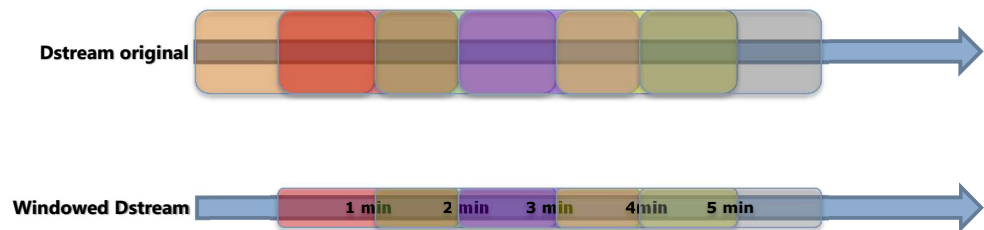
Cas 1 :

Je cherche les données sur les 2 dernières minutes toutes les minutes.

- Window length = 2 minutes
- Sliding interval = 1 minutes



Window et sliding



Une formation
Alphorm.com



Window et sliding

Cas 2 :

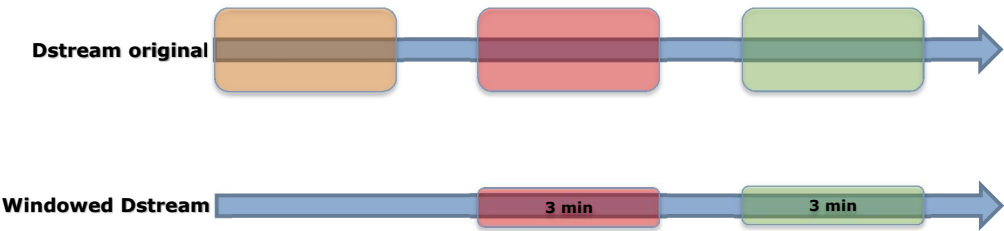
Je cherche les données sur les 2 dernières minutes toutes les 3 minutes.

- Window length = 2 minutes
- Sliding interval = 3 minutes

Une formation
Alphorm.com



Window et sliding



Merci



Une formation
Alphorm.com

DStream socketTextStream



Aimad TAHI



Une formation
Alphorm.com

Plan

Source DStream socketTextStream
Opération en Output DStream



Source Dstream

socketTextStream()

```
var mylines = ssc.socketTextStream('localhost',9999)
```

Une formation
Alphorm.com



Les Opérations en Output DStream

print()

```
Dstream.print()
```

Une formation
Alphorm.com

Merci



DStream textFileStream, foreachRDD



Aimad TAHI



Plan

Sources DStream

Les Opérations en Output DStream

Une formation
Alphorm.com



Sources DStream

textFileStream()

```
var mylines = ssc.textFileStream('hdfs:///data/incoming/')
```

Une formation
Alphorm.com



Les Opérations en Output DStream

print()

```
Dstream.print()
```

saveAsTextFiles()

```
DStream.saveAsTextFiles(prefix, suffix--None)
```

foreachRDD()

```
DStream.foreachRDD(func)
```

Merci



Une formation
Alphorm.com

DStream Sliding Window



Aimad TAHI



Une formation
Alphorm.com

Plan

Sliding Window Operations
Cas de Sliding Window



Sliding Window Operations

Transformation	Fonctionnement
Window (windowLength, slideInterval)	Renvoie un nouveau DStream calculé en fonction des windows du flux DStream source.
countByWindow(windowLength, slideInterval)	Renvoie le nombre d'éléments de la window glissantes dans le flux.
reduceByWindow(func,windowLength,slideInterval)	Renvoie un nouveau flux d'élément unique, créé en agrégeant les éléments du flux sur un intervalle de glissement (sliding) en utilisant func. La fonction doit être associative pour pouvoir être calculée correctement en parallèle.
reduceByKeyAndWindow (func, windowLength, slideInterval, [numTasks])	Lorsqu'il est appelé, un DStream de paires (K, V) renvoie un nouveau DStream de paires (K, V), où les valeurs de chaque clé sont agrégées en utilisant la fonction de réduction donnée (func) sur des batchs dans une fenêtre glissante.



Sliding Window Operations

Transformation	Fonctionnement
countByValueAndWindow(windowLength, slideInterval, [numTasks])	Lorsqu'il est appelé sur un DStream de paires (K, V), il renvoie un DStream avec de nouvelles paires (K, Long) où les valeurs de chaque clé présente la fréquence dans une des windows glissantes.



Cas de Sliding Window

Pratique

Je cherche les données sur les 9 dernières secondes toutes les 3 secondes.

- Window length = 9 secondes
- Sliding interval = 3 secondes



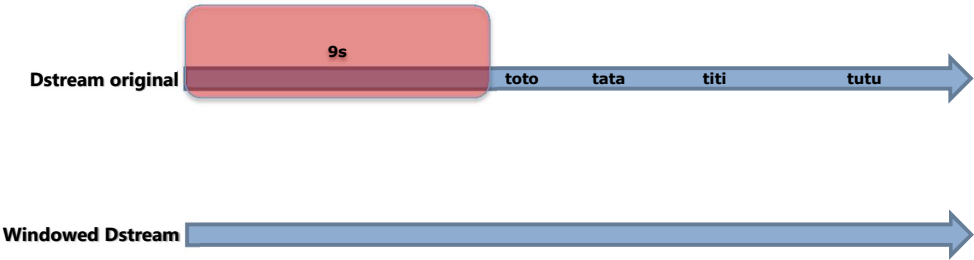
Cas de Sliding Window

Dstream original 

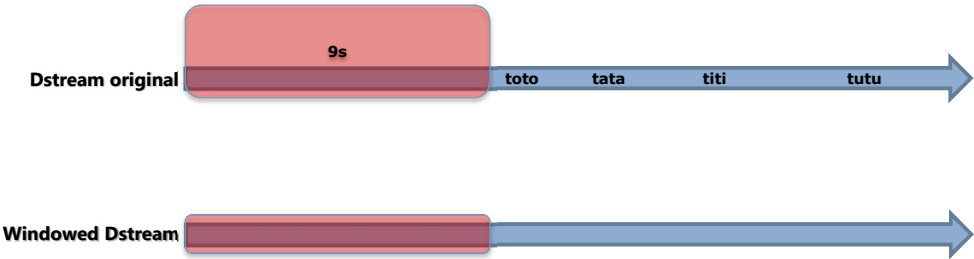
Windowed Dstream 



Cas de Sliding Window

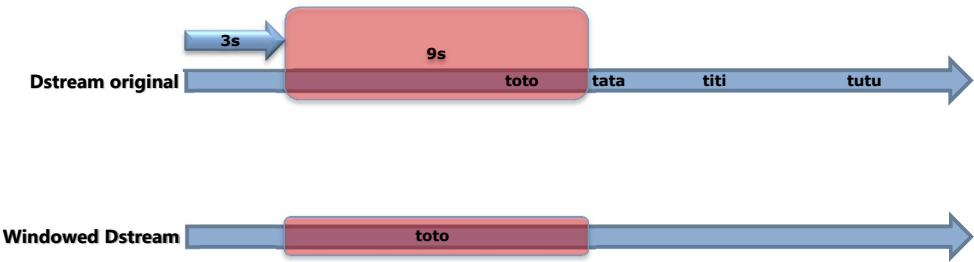


Cas de Sliding Window

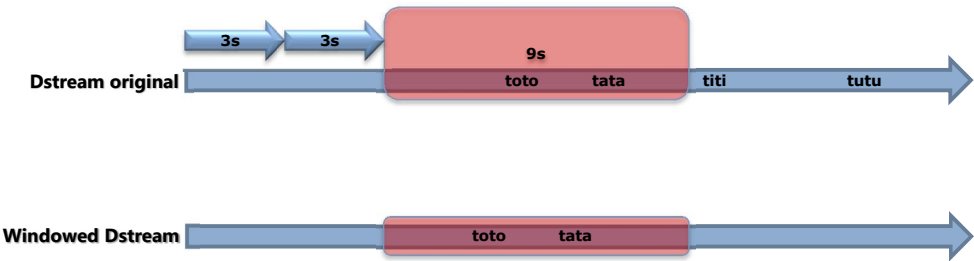




Cas de Sliding Window

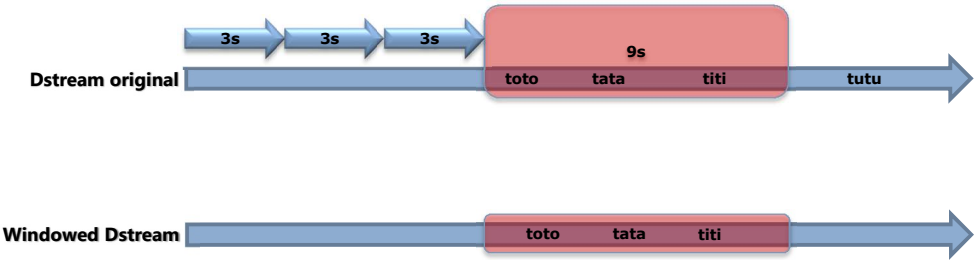


Cas de Sliding Window

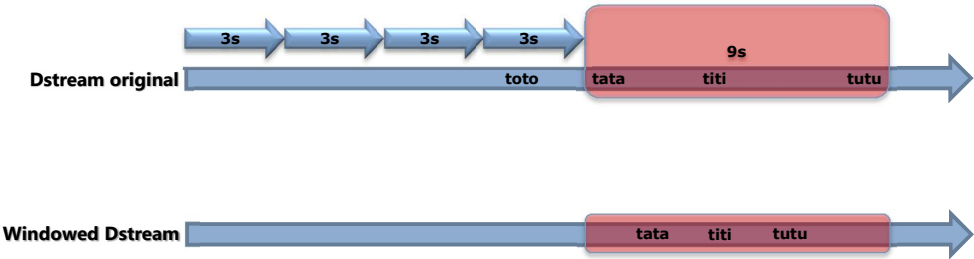




Cas de Sliding Window

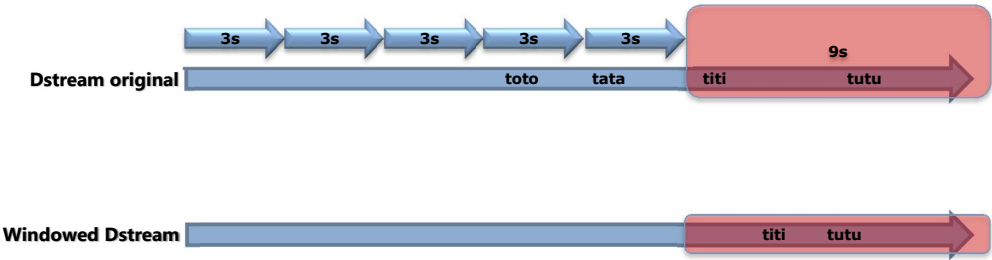


Cas de Sliding Window

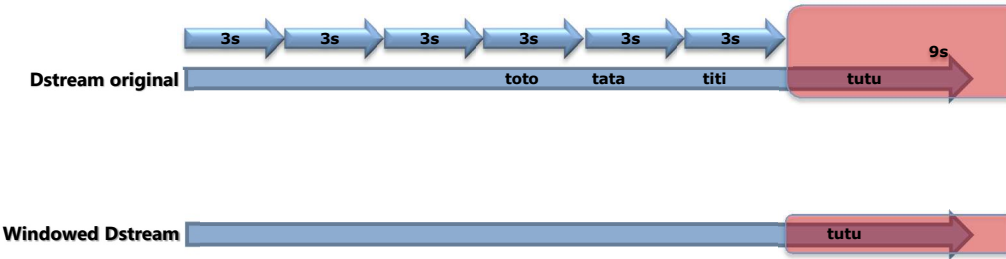




Cas de Sliding Window

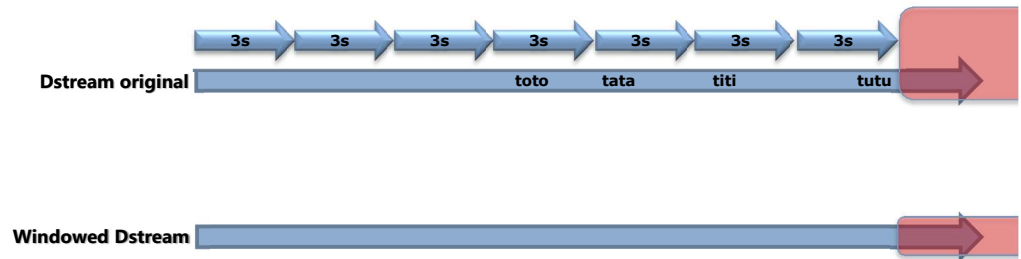


Cas de Sliding Window





Cas de Sliding Window



Une formation
Alphorm.com



Conclusion

Spark Streaming permet de traiter les données au fil de leurs arrivées.



Attention : Si la "window" est trop large, il faut évaluer la pertinence de l'usage du streaming

Une formation
Alphorm.com

Merci



Conclusion



Aimad TAHI



Une formation
Alphorm.com

Bilan

Big DATA

Apache Hadoop

Apache Spark

Scala

RDDs

Spark SQL

Spark Streaming

Merci