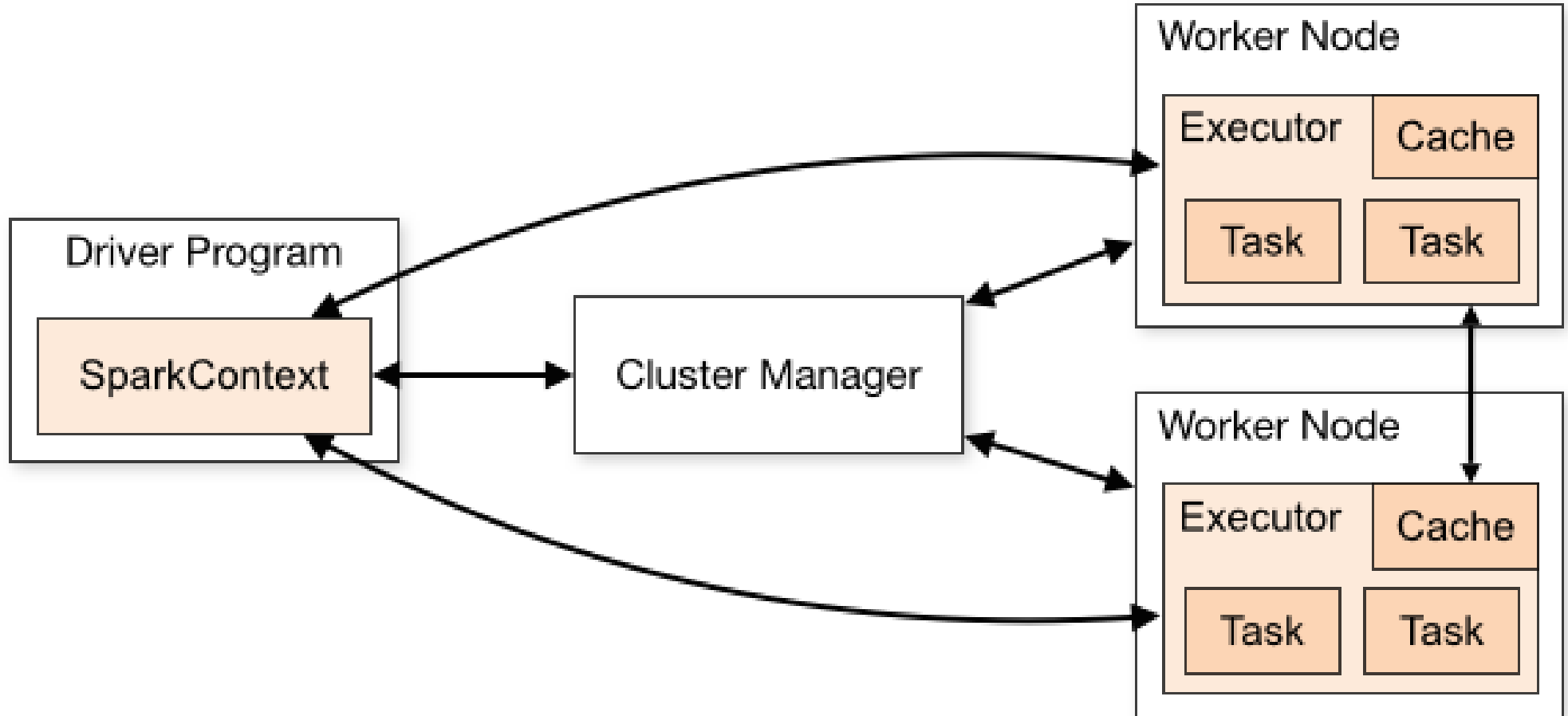


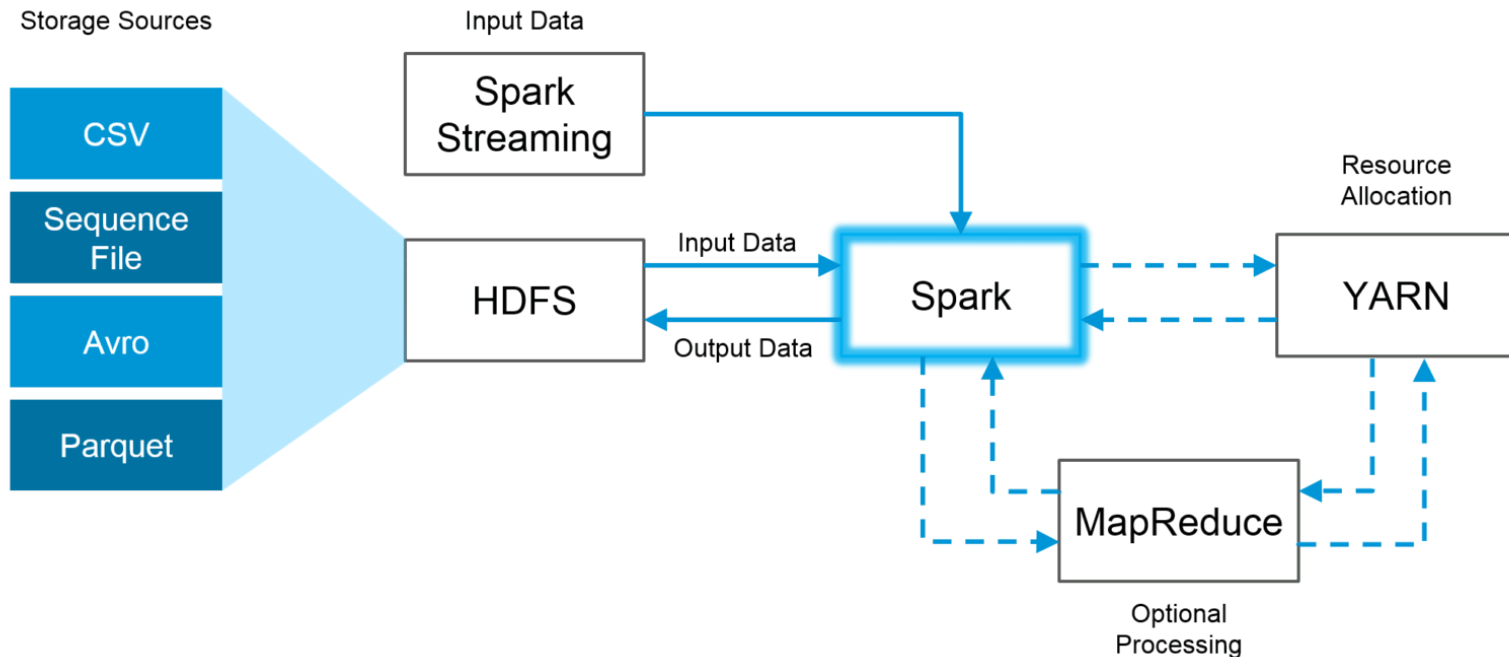


Tour d'horizon : qu'est-ce que Spark et qu'apporte-t-il par rapport aux jobs MapReduce ?

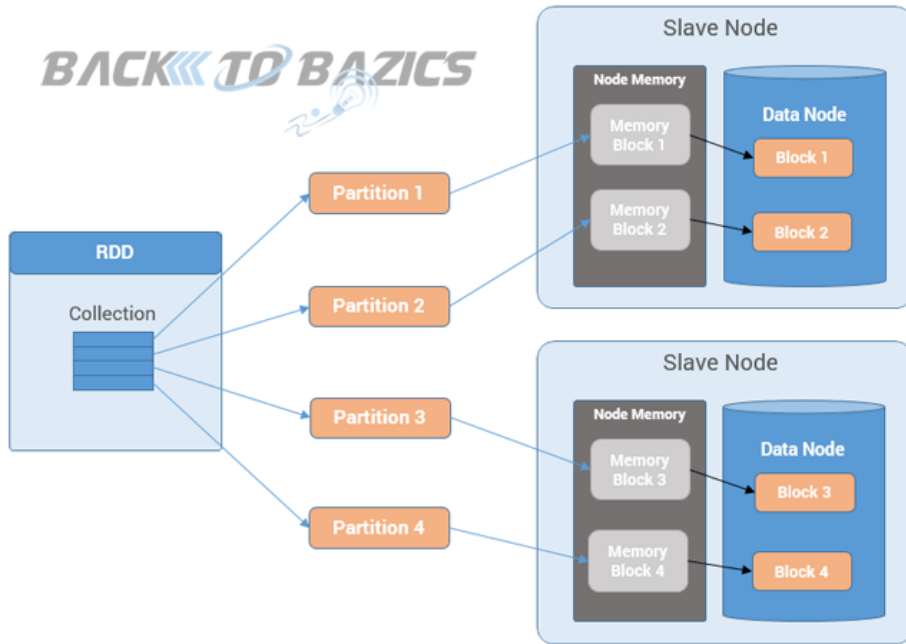
Spark est un framework de calcul distribué



Fonctionnalités similaires à un map reduce



Mais utilisation particulière de la RAM des workers



- Caching
- Temps réel (vs Batch processing)

En résumé



Spark vs Hadoop MapReduce

Factors

Speed

100x times than MapReduce

Faster than traditional system

Written In

Scala

Java

Data Processing

Batch / real-time / iterative /
interactive / graph

Batch processing

Ease of Use

Compact & easier than Hadoop

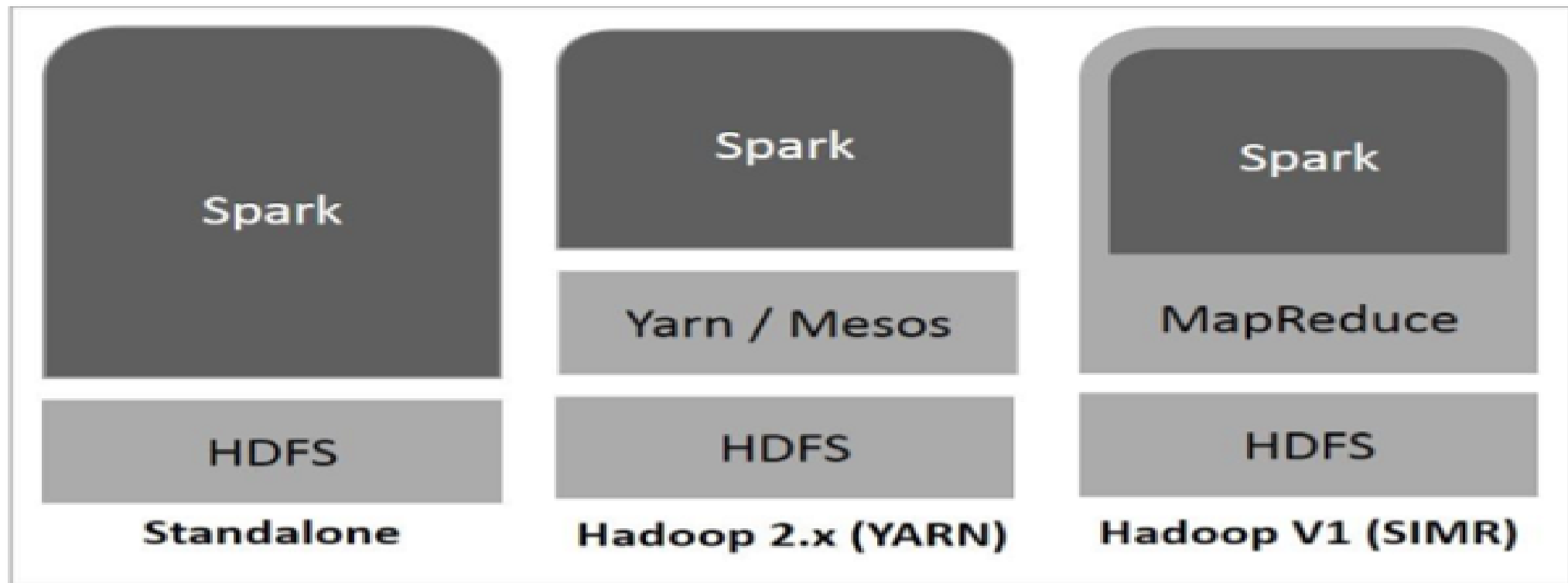
Complex & lengthy

Caching

Caches the data in-memory &
enhances the system performance

Doesn't support caching of data

Intégration dans l'écosystème



Anatomie d'une application Spark

NB : (« application » | Spark) = (« job » | MapReduce)

Une application Spark commence toujours par l'instanciation d'un « **spark context** »

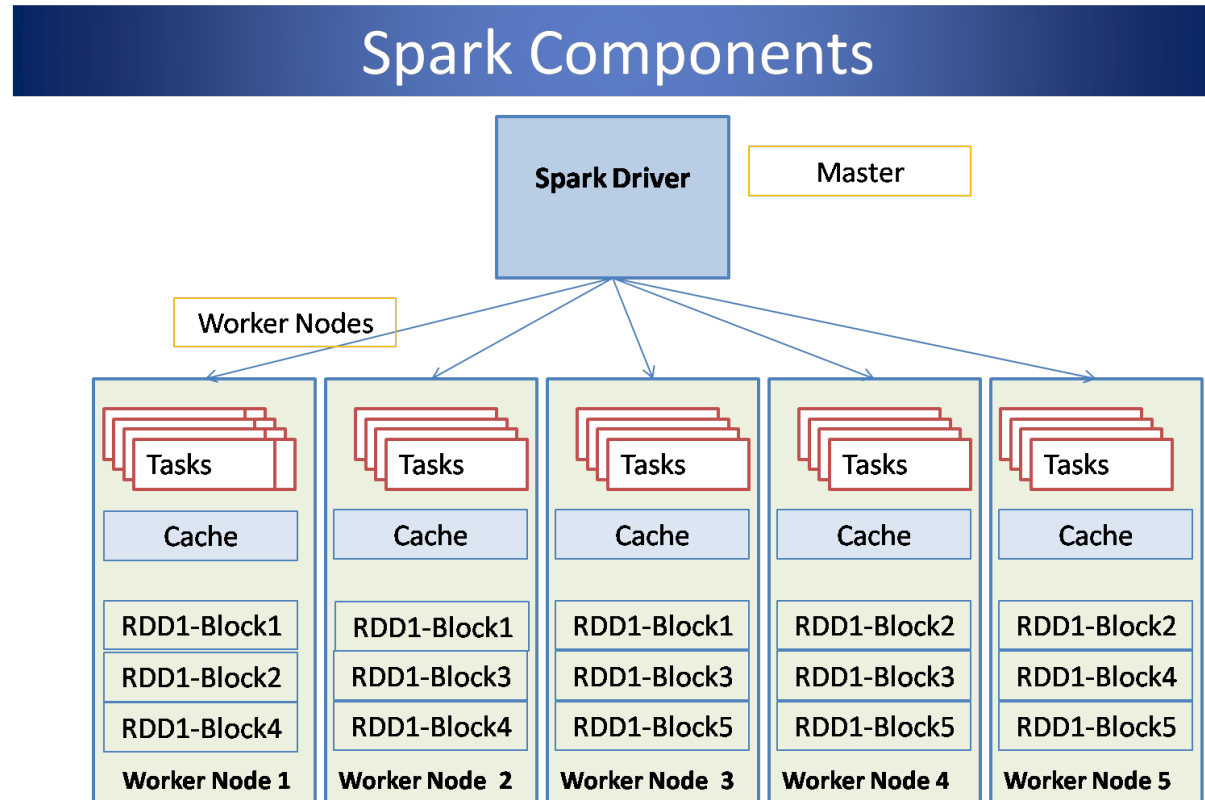
```
from pyspark import SparkContext
```

```
sc = SparkContext("local", "First App")
```

(indique au driver l'existence d'une application pour qu'il lui alloue des workers)

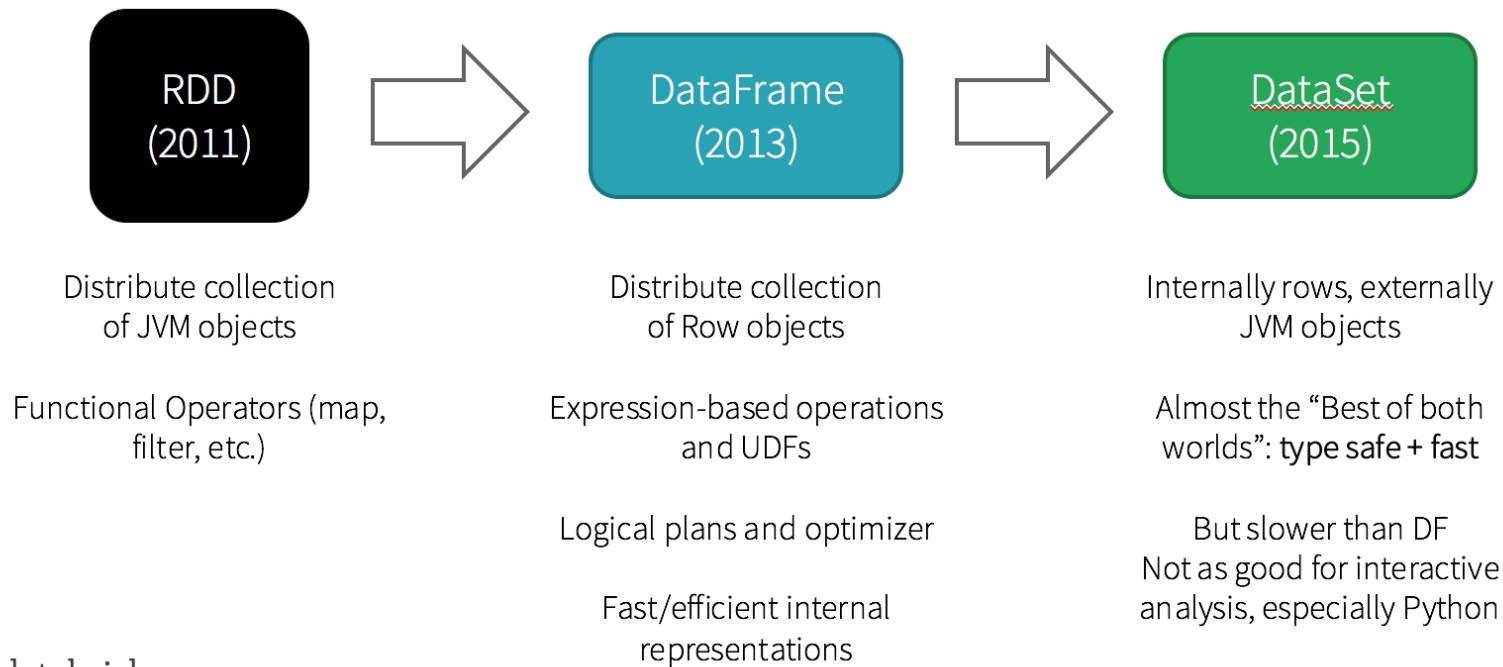
Mais que sont ces RDD (Résilient distributed datasets) ?

En un mot, des collections distribuées sur la ram des workers



Nous rencontrerons d'autres structures de données propres à Spark

History of Spark APIs



Mise en place de l'environnement

Environnement technique

Hortonworks sandbox (virtualbox)

<https://fr.hortonworks.com/tutorial/hortonworks-sandbox-guide/>

Ajout du nom d'hôte

Ajouter la ligne suivante :

```
127.0.0.1 sandbox-hdp.hortonworks.com sandbox-  
hdf.hortonworks.com
```

- Sous linux/MacOSX dans /etc/hosts
- Sous Windows C:\Windows\System32\Drivers\etc

Démarrage des services

Étape 1 se rendre sur **Ambari** (port 8080)

Les credentials sont

- User : **raj_ops**
- Mdp : **raj_ops**

Ambari - Sandbox - Mozilla Firefox

localhost:8080/#/main/services/RANGER/summary

Ambari

Dashboard

Services

- HDFS
- YARN
- MapReduce2
- Tez
- Hive
- HBase
- Pig
- Sqoop
- Oozie
- ZooKeeper
- Storm
- Infra Solr
- Atlas
- Kafka
- Knox
- Ranger

Services / Ranger / Summary

SUMMARY CONFIGS

IC I

Start All

Started RANGER ADMIN

Started RANGER TAGSYNCS

Enabled RANGER HDFS PLUGIN

Enabled RANGER YARN PLUGIN

Enabled RANGER HIVE PLUGIN

Disabled RANGER HBASE PLUGIN

Disabled RANGER STORM PLUGIN

Enabled RANGER ATLAS PLUGIN

Disabled RANGER KAFKA PLUGIN

Enabled RANGER KNOX PLUGIN

Quick Links

Ranger Admin UI

Licensed under the Apache License, Version 2.0.
See third-party tools/resources that Ambari uses and their respective authors

Démarrage des services

Assurez-vous que les services suivants fonctionnent (le démarrage de l'ensemble de ces services est un peu long, c'est tout à fait normal)

- HDFS
- YARN
- HIVE (spark.sql va en avoir besoin)
- SPARK2
- ZEPPELIN NOTEBOOK (un genre de jupyter)

Importation des notebooks dans Zeppelin

The screenshot shows the Zeppelin web interface in a Mozilla Firefox browser. The browser's address bar displays `sandbox-hdp.hortonworks.com:9995/#/`. The Zeppelin interface has a dark blue header with the logo, 'Notebook' dropdown, and 'Job' button. A search bar and 'anonymous' user indicator are on the right. The main content area says 'Welcome to Zeppelin!' and describes it as a web-based notebook. On the left sidebar, under 'Notebook', the 'Import note' link is highlighted with a red arrow. A modal dialog titled 'Import New Note' is open in the center. It has a blue header with a close button. Below the title, it says 'Import As' followed by a text input field labeled 'Insert Note Name'. A warning message states 'JSON file size cannot exceed 1 MB'. At the bottom, there are two large buttons: 'Select JSON File' with a cloud and upload icon, and 'Add from URL' with a chain link icon.

Zeppelin - Mozilla Firefox

Beyond Creation x Hortonworks Sandbox x Ambari - Sandbox x Ambari - Sandbox x Zeppelin x Trending YouTube x w Apache Spark x 219-screen-shot-2 x Apache Spark RDD x e! Top 55 Apache x Résultats Google x

sandbox-hdp.hortonworks.com:9995/#/ 140% spark

Les plus visités LinkedIn Débuter avec Firefox From Google Chrome Entrepreneuriat Informatique Google Agenda Wordreference Google Scholar pistes pro Linguee Dictionnaire a... Mon Drive - Google Drive

Zeppelin Notebook Job Search anonymous

Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analysis. You can make beautiful data-driven, interactive, collaborative notebooks.

Notebook ↻

- Import note ←
- Create new note

Filter

- Getting Started
- Labs
- first_pyspark
- py_spark_dataframe
- R (SparkR)
- Zeppelin Tutorial (Basic Features)

Import New Note

Import As

Insert Note Name

JSON file size cannot exceed 1 MB

Select JSON File

Add from URL

Importation des données dans l'HDFS

The screenshot shows the Ambari web interface in a Mozilla Firefox browser. The page title is "Ambari - Sandbox". The breadcrumb navigation path is "/ > user > raj_ops". A yellow box indicates "Total: 1 files or folders". A table displays the contents of the directory:

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Encrypted
youtube	--	2019-08-20 18:06	raj_ops	hdfs	drwxr-xr-x	No

A "Views" dropdown menu is open, showing the following options:

- YARN Queue Manager
- Files View (highlighted with a red circle)
- Workflow Manager

The "Files View" option is the selected view for the current directory.

Troubleshooting

Erreur Zeppelin : interpreter not found

The screenshot shows the Zeppelin Notebook interface in a Mozilla Firefox browser. The browser's address bar displays the URL `sandbox-hdp.hortonworks.com:9995/#/notebook/2EMDKC514`. The Zeppelin interface includes a top navigation bar with the Zeppelin logo, 'Notebook' and 'Job' tabs, a search bar, and a user profile dropdown labeled 'anonymous'. Below this, the notebook title 'first_pyspark' is shown with various action icons. A 'Settings' section is visible, containing an 'Interpreter binding' subsection. This section explains how to bind an interpreter and provides a 'Restart' button. Below the button is a list of interpreters: 'spark2' (with a default icon), 'angular', 'jdbc', 'livy2', and 'md'. At the bottom of the notebook area, there are 'Save' and 'Cancel' buttons. The bottom of the image shows the start of a new code block titled 'Prise en main de PySpark'.

first_pyspark - Zeppelin - Mozilla Firefox

sandbox-hdp.hortonworks.com:9995/#/notebook/2EMDKC514

Zeppelin Notebook Job

Search anonymous

first_pyspark

Settings

Interpreter binding

Bind interpreter for this note. Click to Bind/Unbind interpreter. Drag and drop to reorder interpreters. The first interpreter on the list becomes default. To create/remove interpreters, go to [Interpreter](#) menu.

Restart

spark2 %spark2 (default), %sql, %dep, %pyspark, %ipyspark, %r

angular %angular

jdbc %jdbc

livy2 %livy2, %livy2.sql, %livy2.pyspark, %livy2.sparkr, %livy2.shared

md %md

Save Cancel

Prise en main de PySpark

FINISHED

Les jobs ne se lancent pas => YARN

Il peut être intéressant de regarder si **trop de jobs** ne sont pas en cours sur la plateforme et les KILL au besoin

<http://sandbox-hdp.hortonworks.com:8088>

The screenshot displays the Apache Hadoop YARN web interface, specifically the 'Applications' tab. The top navigation bar includes links for Cluster Overview, Queues, Applications (selected), Services, Flow Activity, Nodes, and Tools. The user is logged in as 'di'. The main content area shows a list of applications with columns for Application ID, Application Type, Application Name, User, State, Queue, and Progress. A sidebar on the left allows filtering by User (zeppelin, hive) and State (FINISHED, KILLED, FAILED, RUNNING). The 'zeppelin' user has 12 applications, and the 'hive' user has 9. The 'RUNNING' state has 1 application. The 'Application ID' column shows 'application_1566398080487_00...'. The 'Application Type' is 'SPARK'. The 'Application Name' is 'Zeppelin'. The 'User' is 'zeppe...'. The 'State' is 'Ru...'. The 'Queue' is 'default'. The 'Progress' is 10%.

User (2)	Application ID	Application T...	Application ...	User	State	Queue	Progress
<input checked="" type="checkbox"/> zeppelin 12	application_1566398080487_00...	SPARK	Zeppelin	zeppe...	● Ru...	default	10%
<input checked="" type="checkbox"/> hive 9							

More

State (4) All

- ☐ FINISHED 8
- ☐ KILLED 7
- ☐ FAILED 5
- ☒ RUNNING 1

More

Apply Clear