cidacs
Centro de Integração de Dados
e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Descrição do Cluster de Testes

George C. G. Barbosa

7 de março de 2019

Tópicos

- Fontes utilizadas
- Ferramentas
- Testes (provisórios)

Cluster Hadoop e Cluster Spark

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

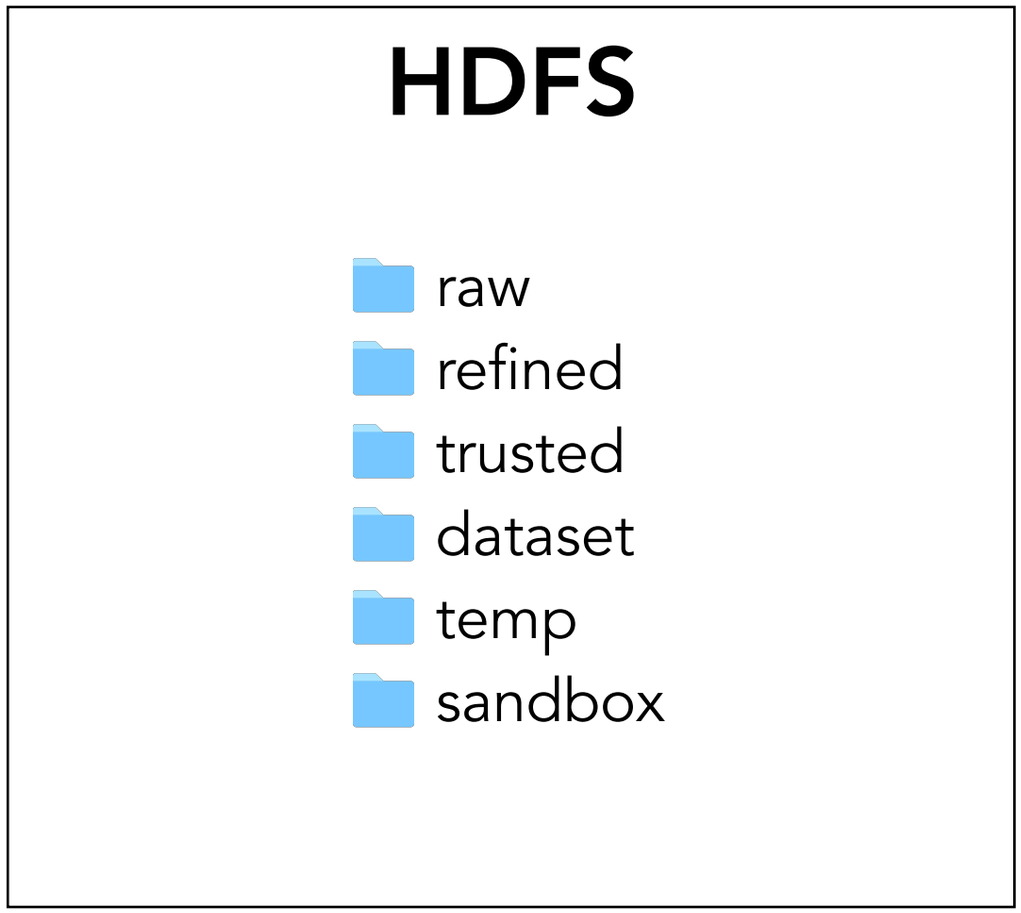
Fontes

- Setting up a scalable data exploration environment with Spark and Jupyter Lab
- Building our data science platform with Spark and Jupyter
- Hadoop documentation 2.9.2
- Spark documentation 2.4.0

Ferramentas

- Hadoop
- Jupyter
- Spark
- Livy

Hadoop



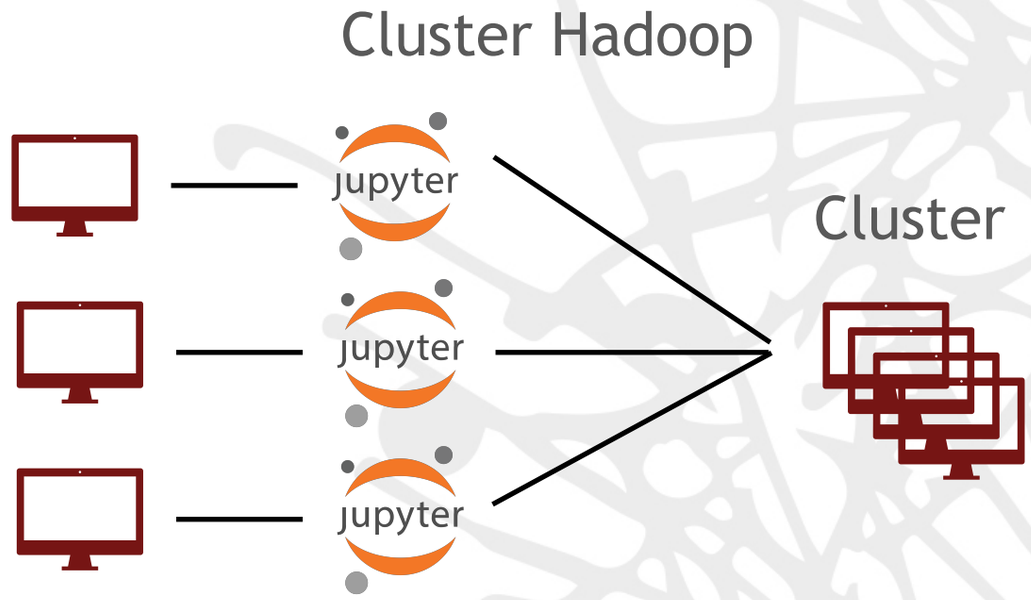
- Data lake
- Tolerância a falhas
- Acesso distribuído aos arquivos

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

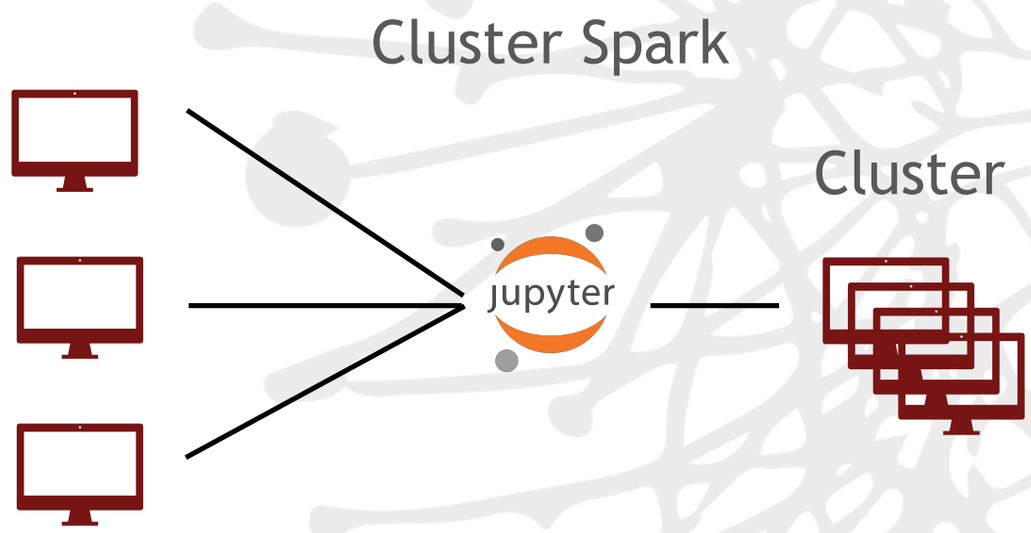
Jupyter

- JupiterHub
- Spark Magic
- Custom Kernel



Cluster Hadoop

- SSH
- Seleção manual de parâmetros
- Cada usuário executa seu próprio cliente



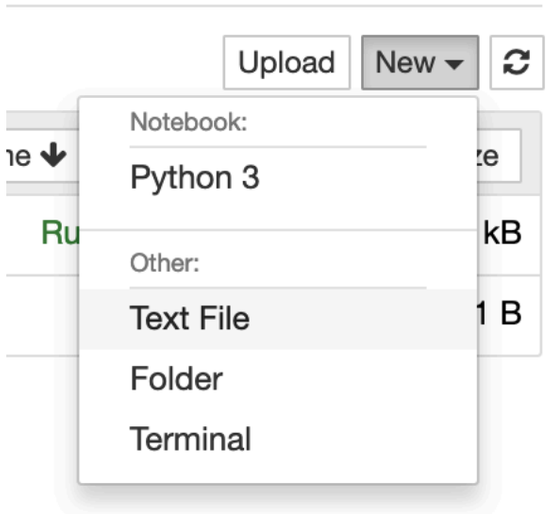
cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Livy

- Spark Magic
- Kernels
- Sessões de Longa duração
- Falta do Autocompletar

Solucão: Custom Kernel



cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

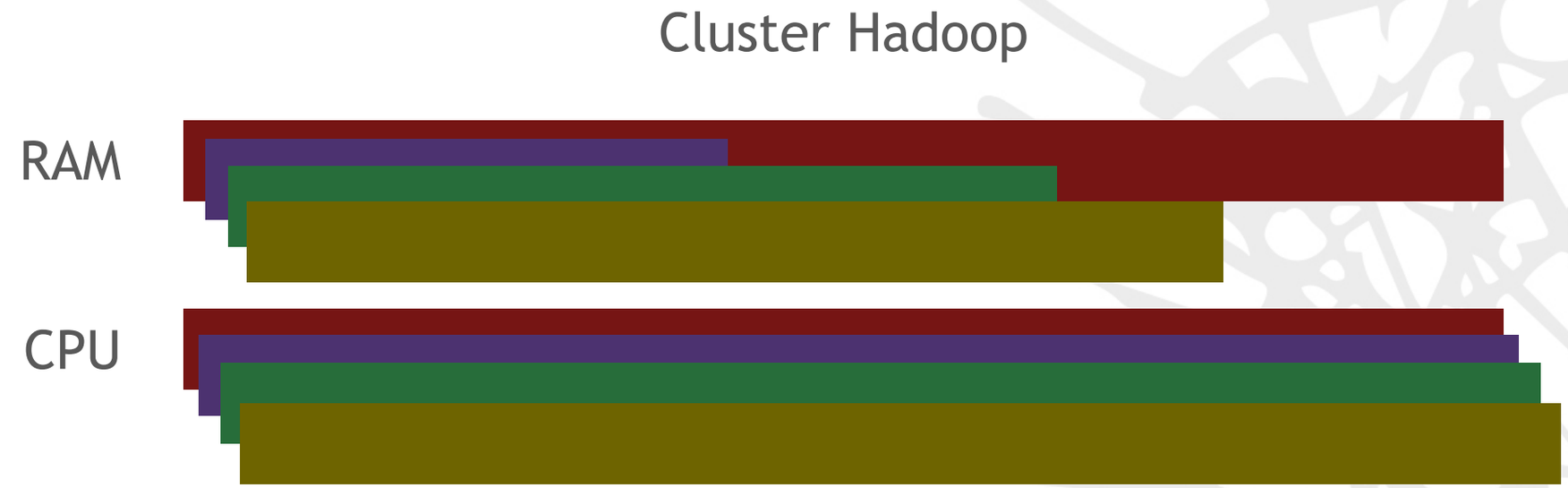
Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Spark

- Standalone Cluster
- Scheduling
  - FIFO
  - Dynamic Resource Allocation

Spark

- Sem fila?



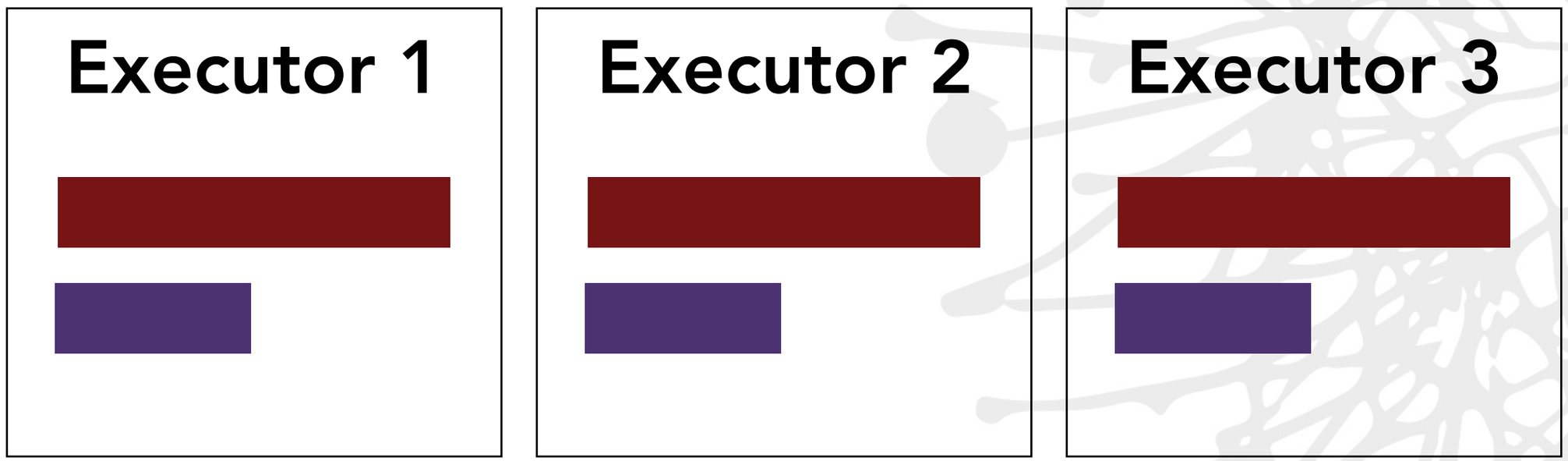
Spark

- Gerenciamento por executor
- Fila de processos

Cluster Spark

Spark

- Gerenciamento por executor
- Fila de processos



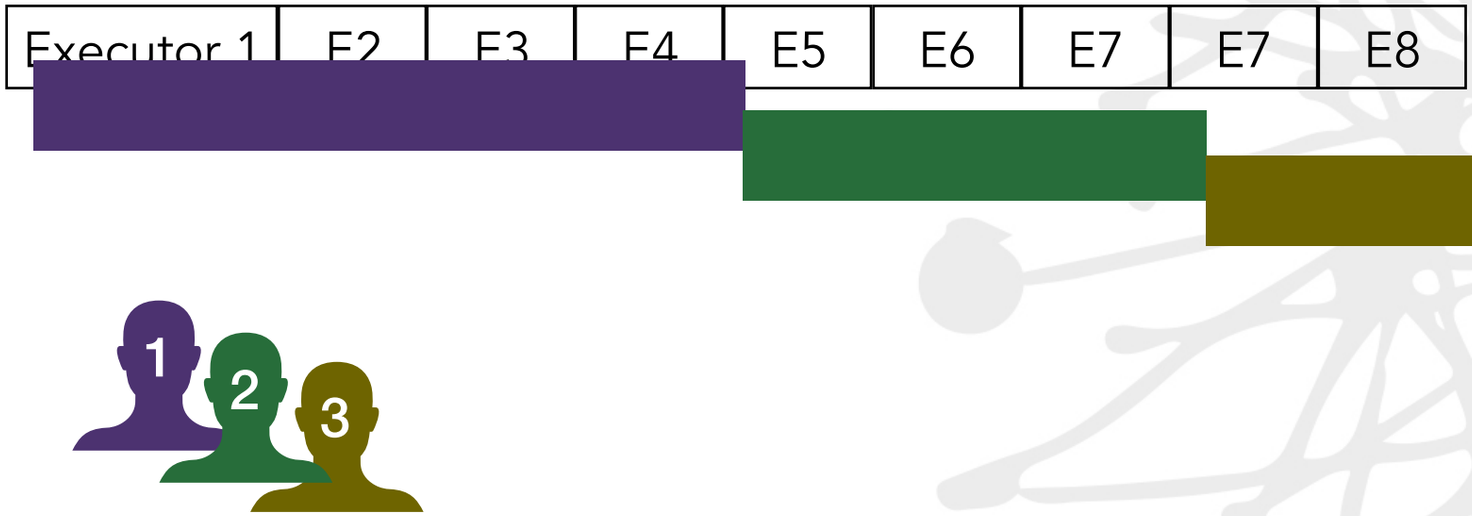
cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Spark

- Gerenciamento por executor
- Fila de processos

Cluster Spark



Testes

- Processo de longa duração
- Compartilhamento de recursos
  - 40+ Jobs a partir de 2 usuários diferentes

Application ID ▲	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20190225122038-0011 (kill)	pyspark-shell	8	2.0 GB	2019/02/25 12:20:38	cluster	RUNNING	50.2 h

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz
Instituto Gonçalo Moniz

Documentação

- https://github.com/cidacslab/cidacs-cluster/blob/master/documentation/tutorial.md