Databricks

## O que é?

Desenvolvido pelos mesmos criadores do Apache Spark, o Databricks é uma plataforma de análise de dados unificada para engenharia de dados, machine learning e ciência de dados colaborativa. Se organiza através de workspaces, que são ambientes de software como serviço (SaaS) utilizados para acessar suas funcionalidades e objetos, como notebooks, em pastas e recursos computacionais, como clusters.

## Por que utilizar Databricks?

### Workspace colaborativo

Uma primeira razão para utilizar Databricks está em seus workspaces colaborativos, onde usuários diferentes podem compartilhar seus notebooks ou até mesmo trabalhar em um de forma simultânea; experiência similar a do Google Docs ou Word Online.

### Para Cientistas, Analistas e Engenheiros de Dados

Databricks é uma plataforma poderosa que pode ser utilizada de maneira agradável por diferentes profissionais. Ela permite que cientistas de dados manipulem dados e criem modelos para Machine Learning; que analistas usem tabelas estruturadas, arquivos semi-estruturados e se conectem a ferramentas de visualização nativamente, e; que engenheiros de dados extraiam, transformem e carreguem dados em diferentes ambientes.

### Pluralidade de Linguagens

Os notebooks Databricks, similares ao do Jupyter, não se limitam a uma só linguagem de programação: neles é possível utilizar Python, SQL, R e Scala. Sendo possível, inclusive, mesclar em um mesmo notebook células em linguagens diferentes e combiná-las posteriormente, como por exemplo criando uma função em Python em uma primeira e a utilizando em um código SQL em uma segunda.

A pluralidade de linguagens resolve diversos problemas que times enfrentam ao definir arquitetura e escolher produtos. Se parte dos cientistas for mais familiarizada com Python e outra com R, tudo certo; se alguns engenheiros preferem utilizar Spark em Scala e outros em Python, tudo certo também. Evita-se assim, uma necessidade de adaptar e treinar elementos do time a novas skills.

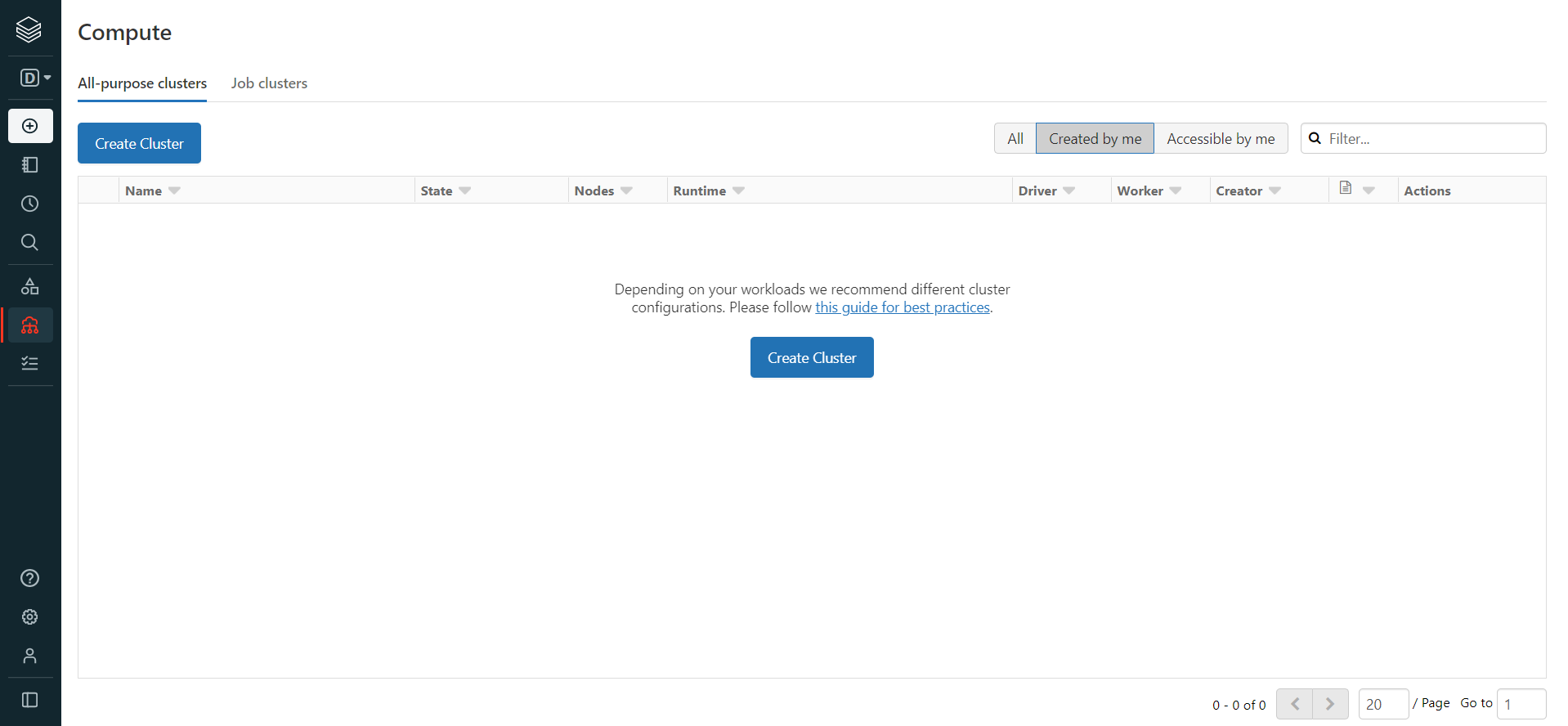
### Multi-cloud

Databricks é uma ferramenta que pode ser considerada agnóstica por ser multicloud, em outras palavras, se conecta e se complementa muito bem a diferentes plataformas de computação em nuvem, como Amazon Web Services (AWS), Google Cloud Plataform (GCP) e Microsoft Azure. Sendo assim, independente da plataforma escolhida para sediar a arquitetura de uma solução, é possível encaixar produtos do Databricks em uma ou várias etapas.

## Versão Community

O Databricks possui uma versão Community para que estudantes e interessados na plataforma possam degustar e praticar em uma versão gratuita, ainda que com limitações. Outro aspecto importante é que diferente de suas parceiras, AWS Azure e GCP, a versão community é totalmente gratuita. Aquilo que está disponível apenas em versões pagas, simplesmente não funciona. Ou seja, não é preciso inserir um cartão de crédito e fazer um rígido controle dos produtos utilizados para evitar gastos inesperados enquanto está apenas experimentando.

### Cluster

A tela a seguir é acessada através do menu "Compute" à esquerda. Aqui podemos criar um cluster no botão "Create Cluster", dando um nome qualquer a ele e mantendo as próximas configurações padrões. Caso o projeto que venha a desenvolver requeira uma versão específica do Spark ou Scala, basta alterar durante a criação. Quanto ao poder computacional (quantidade de nós, memória, etc.), não podemos alterá-lo na versão community. 

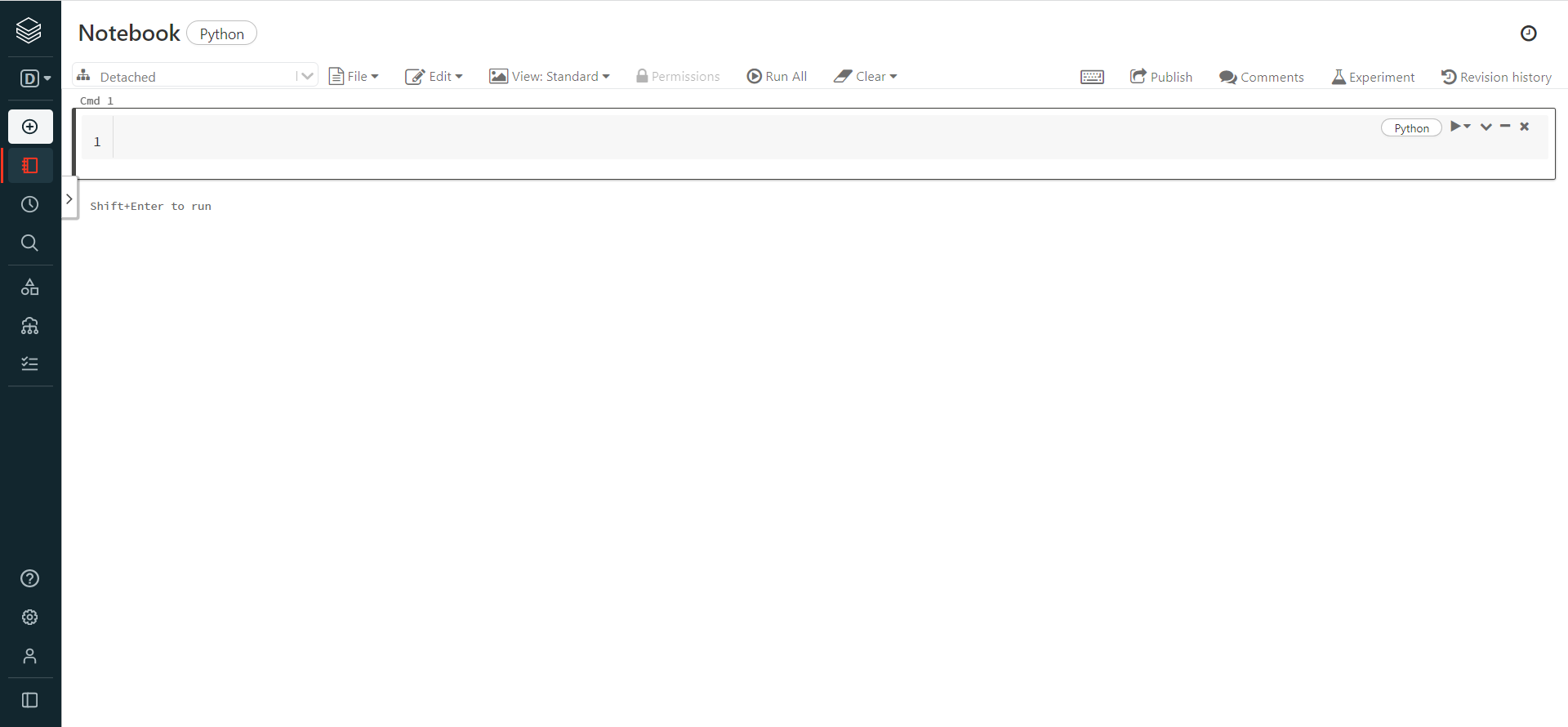
## Notebook

"Um notebook é uma interface baseada na Web para um documento que contém código executável, visualizações e texto de narração."

Composto por células capazes de executar códigos em diferentes linguagens, é em um notebook que somos capazes de manipular arquivos, criar funções e tabelas, etc.

Caso tenha utilizado Jupyter anteriormente, saiba que o conceito é o mesmo, assim como a extensão do arquivo .ipynb.

Na imagem abaixo temos um exemplo de notebook, em branco, criado em um diretório qualquer no Workspace (no menu à esquerda clicar em workspace e, no menu que se abrir, selecionar o usuário - em um ambiente colaborativo é possível que outros usuários, além do seu, estejam disponíveis. Na área em branco, clicar no botão direito -> Create -> Notebook). Note que, ao lado de seu nome, temos a sua linguagem default (Python), o que quer dizer que por padrão as células esperarão códigos em Python. Abaixo do nome do notebook, selecionamos o cluster em que os códigos serão executados (lembrando que a versão community do Databricks nos disponibiliza um gratuitamente). Podemos criar quantas células julgarmos necessárias e, para mudar a linguagem de uma individualmente basta selecionar no menu à direita da célula ou em sua primeira linha colocar '#' seguido do nome da linguagem.

Um outro recurso interessante para estudos futuros é acionado no canto superior direito do notebook e diz respeito a versionamento, pois o Databricks permite integração com git em versões pagas (Settings -> User Settings -> Git Integration). 

### Limitações

Algumas limitações que você encontrará ao utilizar a versão community do Databricks que valem ser citadas:

* O cluster disponibilizado, de forma gratuita, é desligado após duas horas de inatividade e não pode ser reinicializado. Com isso, embora arquivos no DBFS e notebooks não sejam perdidos, metadados como tabelas criadas serão.
* Conexões com serviços de armazenamento (mount), como S3 e Blob Storage, utilizando chaves de segurança não são permitidas.

## Indicações e Bibliografia

[Databricks Community](https://community.cloud.databricks.com/)

[Documentação oficial](https://docs.databricks.com/)

[Databricks getting started](https://docs.databricks.com/getting-started/quick-start.html)