



Apache Zeppelin: O Laboratório Open Source para Cientistas de Dados

Marcio Junior Vieira
CEO & Data Scientist, Ambiente Livre
Pesquisador da UFG.

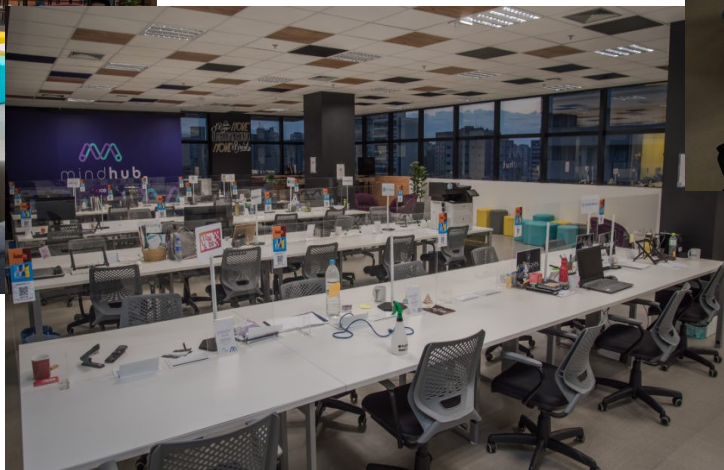
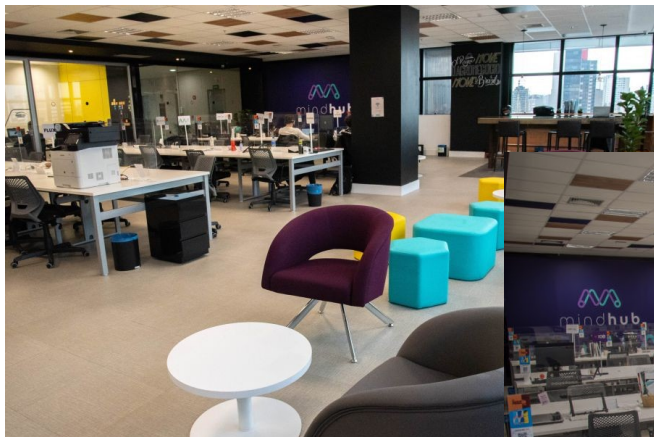
Mini-CV

- 25 anos de experiência em TI, vivência em desenvolvimento e análise de sistemas de gestão empresarial e ciência de dados.
- CEO da Ambiente Livre atuando como Cientista de Dados, Engenheiro de Dados e Arquiteto de Software.
- Professor dos MBAs em Big Data & Data Science, Inteligência Artificial e Business Intelligence e Analytics da Universidade Positivo.
- Professor do MBA Artificial Intelligence e Machine Learning da FIAP.
- Pesquisador do Laboratório de Tecnologias para Tomada de Decisão da Universidade de Brasília (Unb/Latitude).
- Trabalhando com Free Software e Open Source desde 2000 com serviços de consultoria e treinamento.
- Graduado em Tecnologia em Informática(2004) e pós-graduado em Software Livre(2005) ambos pela UFPR.
- Palestrante FLOSS em: FISL, TDC, Latinoware, Campus Party, Pentaho Day, Ticnova, PgDay e FTSL.
- Organizador Geral: Pentaho Day 2017, 2015, 2019 e apoio nas ed. 2013 e 2014.
- Data Scientist, instrutor e consultor de Big Data e Data Science com tecnologias abertas.
- Ajudou a capacitar equipes de Big Data na IBM, Accenture, Tivit, Serpro, Natura, MP, Netshoes, Embraer entre outras.
- Especialista em implantação e customização de Big Data com Hadoop, Spark, Pentaho, Cassandra e MongoDB.
- Contribuidor de projetos internacionais, tais como Pentaho, LimeSurvey, SuiteCRM e Camunda.
- Especialista em implantação e customização de ECM com Alfresco e BPM com Activiti, Flowable e Camunda.
- Certificado (Certified Pentaho Solutions) pela Hitachi Vantara (Pentaho).
- Membro da The Order Of de Bee (comunidade Alfresco para desenvolver o ecossistema Alfresco independente)
- Trabalha profissionalmente com Apache Zeppelin desde 2019.



Open Software for Business

- Fundada em 2004 com foco em consultoria com FLOSS.
- Experts em 34 soluções para geração de negócios com Software Livre/Código Aberto.
- Atualmente estamos sediados no Hub de Inovação Mindhub em Curitiba (FAE).







Nosso Ecossistema de Serviços

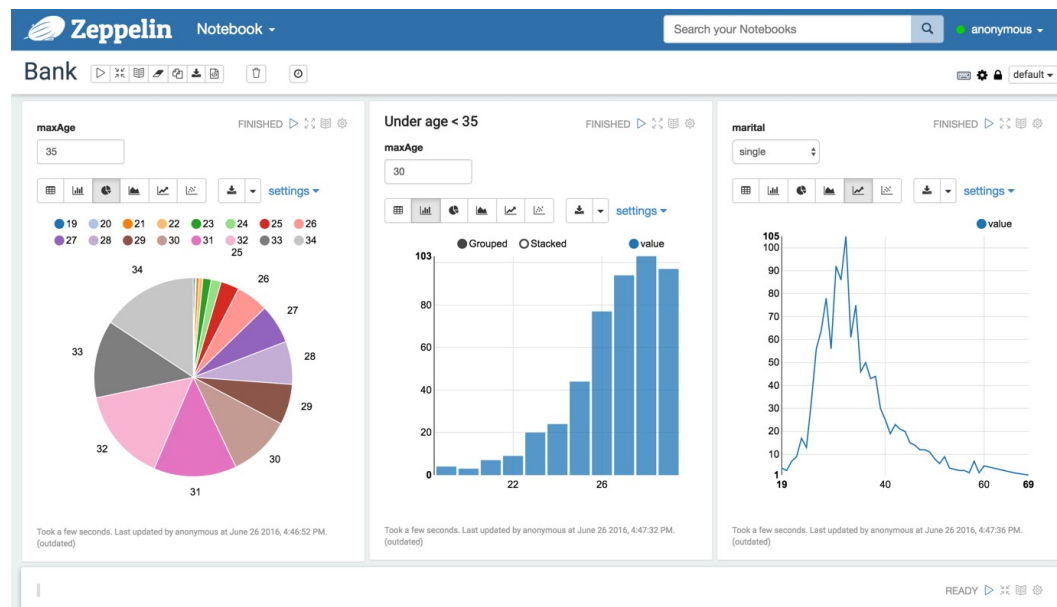
Big Data e Data Science	CRM e CMS	ECM e BPM	Business Intelligence
Análise de Dados da IoT Análise Preditiva Processamento Distribuído Banco de Dados Colunares	Marketing e Vendas Fidelização SAC e Pós-vendas Portais de Conteúdo	Gestão de Documentos Gerenciamento de Mídias Processo de Negócio BPMN e BPMS	Painéis de Indicadores Cubos de Análise Relatórios Gerenciais Tomada de Decisão
Big Data & Data Lake Big Data Analytics Machine Learning	Customer Relationship Management Content Management System Pesquisa de Mercado & SLA	Enterprise Content Management Records Management Business Process Management	Business Intelligence & Analytics Dashboards e OLAP Data Integration & Data Mining
Consultoria Treinamento Projeto	Consultoria Treinamento Projeto	Consultoria Treinamento Projeto	Consultoria Treinamento Projetos



Caderno Multiuso Web

- Uma ferramenta que possibilita análise interativa de dados e documentos
- Colaborativo com SQL, Scala entre outros.
- 100% Open Source.
- Mantido pela Fundação Apache.
- Controle de Acessos.

-  Data Ingestion
-  Data Discovery
-  Data Analytics
-  Data Visualization & Collaboration



Definição

- 491 Projetos Open Source.
- +7000 Committers, e com uma média de 50 novos mensais... Seja um!
- Data Science = Apache = Open Source
- **Apache é líder em Big Data e Data Science!**
- ~49 projetos da linha “Big Data” incluindo “Apache Hadoop” e “Spark”
- ~25 projetos de database incluindo “Apache Cassandra”



Patrocinadores da Apache Software Foundation.

PLATINUM SPONSORS:



LeaseWeb



Facebook



Amazon Web Services



Pineapple Fund



Verizon Media



Tencent



Google



Huawei



Comcast

GOLD SPONSORS:

Anonymous



Baidu



Bloomberg



Cloudera



Handshake



IBM



Union Investment



Workday

SILVER SPONSORS:



Aetna



Alibaba Cloud Computing



Budget Direct



Capital One



Cerner



Inspur



Red Hat, Inc.



Target



Conceito de Interpretador

- O conceito de intérprete do Apache Zeppelin permite que qualquer backend de processamento de dados / linguagem seja conectado ao Zeppelin.
- Atualmente, o Apache Zeppelin suporta muitos intérpretes, como Apache Spark, Python, JDBC, Markdown e Shell.



ALLUXIO



Google BigQuery



cassandra



elasticsearch



Flink



APACHE
GEODE

APACHE
HBASE



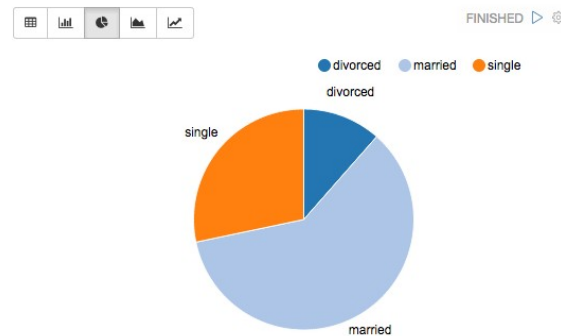
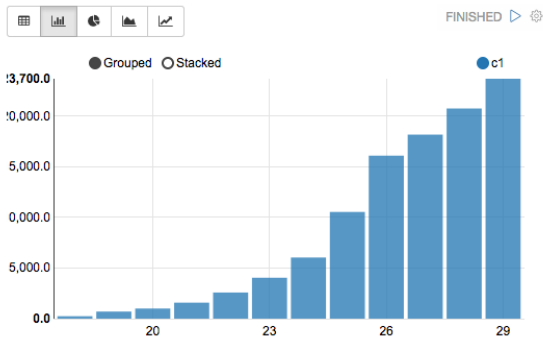
Spark

- Especialmente, o Apache Zeppelin fornece integração completa do Apache Spark. Você não precisa construir um módulo separado, plugin ou biblioteca para isso.
- Injeção automática do SparkContext e SQLContext.
- Carregamento da dependência jar do tempo de execução do sistema de arquivos local ou do repositório maven.
- Cancelamento de job e exibição do seu progresso.



Data Visualization

- Alguns gráficos básicos já estão incluídos no Apache Zeppelin.
- As visualizações não estão limitadas à consulta do Spark SQL, qualquer saída de qualquer backend de linguagem pode ser reconhecida e visualizada.



Data Visualization

- O Apache Zeppelin agrega valores e os exibe no gráfico dinâmico com simples arrastar e soltar.
- Você pode criar facilmente um gráfico com vários valores agregados, incluindo soma, contagem, média, mínimo, máximo.



- O Apache Zeppelin pode criar dinamicamente alguns formulários de entrada no seu bloco de notas.

```
%md Hello ${name=sun}
```

name

Hello moon

```
%spark  
println("Today is "+z.select("day", Seq(("Monday", "1"),  
    ("Tuesday", "2"),  
    ("Wednesday", "3"),  
    ("Thursday", "4"),  
    ("Friday", "5"),  
    ("Saturday", "6"),  
    ("Sunday", "7"))))
```

day

Today is Friday



- Cada Note pode ser considerada um projeto ou um escopo que deseja documentar pelo Apache Zeppelin.
- Um note pode ter diversos parágrafos dentro dele.

Notebook

 Import note

 Create new note

 Filter

 Curso Zeppelin

 Getting Started

 Labs

 R (SparkR)

 Zeppelin Tutorial (Basic Features)



- Parágrafos são as subdivisões que temos de nossas anotações no Apache Zeppelin.
- Cada paragrafo pode ter um interpreter diferente assim com configurações de apresentação.

Ubuntu Software

Some initial delay to be expected...

FINISHED ▶ 🔍 📖 ⚙️

Note: The first time you run `spark.version` in the paragraph below, several services will initialize in the background. This may take **1~2 min** so please **be patient**. Afterwards, each paragraph should run much more quickly since all the services will already be running.

Took 0 sec. Last updated by admin at February 22 2017, 12:25:34 PM. (outdated)

Verify Spark Version (should be 2.x)

FINISHED ▶ 🔍 📖 ⚙️

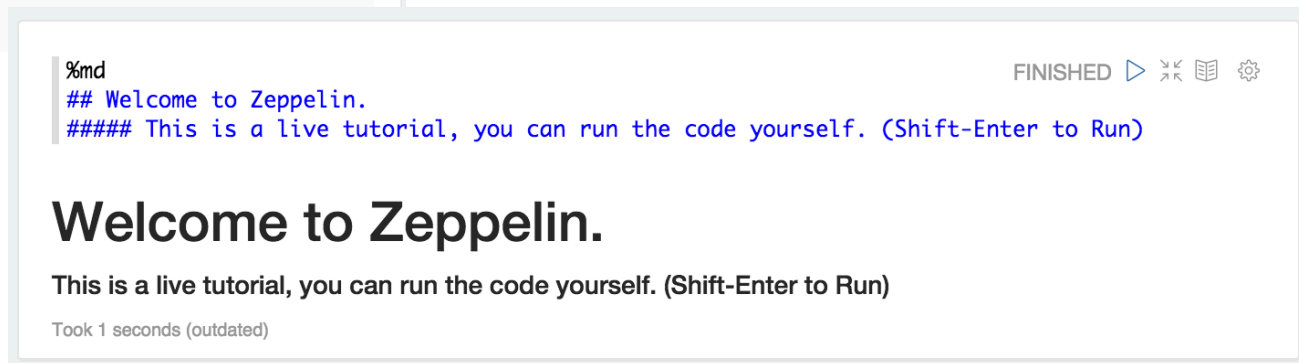
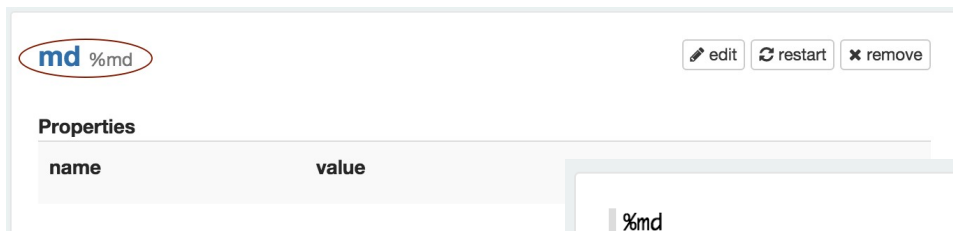
```
%spark2.spark  
  
spark.version
```

Took 21 sec. Last updated by admin at February 17 2017, 3:03:05 AM. (outdated)



Markdown Interpreter - %md

- O Markdown é uma sintaxe de formatação de texto simples projetada para ser convertida em HTML. Zeppelin usa markdown4j.
- No bloco de notas do Zeppelin, você pode usar **%md** no início de um parágrafo para invocar o interpretador Markdown para gerar html estático a partir do texto simples Markdown.



JDBC Interpreter - %jdbc

- O interpretador JDBC permite criar uma conexão JDBC para qualquer fonte de dados sem problemas.
- Inserções, Atualizações e Upserts são aplicados imediatamente após a execução de cada instrução.



Interpreters

Manage interpreters settings. You can create / edit / remove settings. Note can bind / unbind these interpreter settings.

 + Create

Create new interpreter

Interpreter Name

mysql

Interpreter group

jdbc







Option

shared  Interpreter for note

☐ Connect to existing process

☐ Set permission

Properties

name	value	action
common.max_count	1000	
default.driver	org.mysql.jdbc.Driver	
default.password	mysql_password	
default.url	jdbc:mysql://localhost:3306/	
default.user	mysql_username	
zeppelin.jdbc.auth.type		



- O Hive Interpreter trabalha similar ao JDBC Interpreter.
- Você pode usar o Hive Interpreter usando o JDBC Interpreter com a mesma funcionalidade.

```
%hive  
select * from my_table;
```



Shell Interpreter - %sh

- O interpretador de shell usa o Apache Commons Exec para executar processos externos. No bloco de notas do Zeppelin.
- Você pode usar **%sh** no início de um parágrafo para chamar o shell do sistema e executar comandos.

```
%sh
```

```
ls -lh
```

```
total 243K
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 13:09 _tools
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:52 alluxio
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:50 angular
drwxr-xr-x 1 FVALERI 1049089  0 Jun 22 21:07 bin
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 cassandra
drwxr-xr-x 1 FVALERI 1049089  0 Jun 22 21:07 conf
-rw-r--r-- 1 FVALERI 1049089 11K Jun 13 13:09 CONTRIBUTING.md
-rw-r--r-- 1 FVALERI 1049089 777 Jun 13 17:42 derby.log
drwxr-xr-x 1 FVALERI 1049089  0 Jun 22 21:07 dev
-rw-r--r-- 1 FVALERI 1049089 542 Jun 13 13:09 DISCLAIMER
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 13:09 docs
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:52 elasticsearch
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 file
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 flink
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:38 geode
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 hbase
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 hive
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:51 ignite
drwxr-xr-x 1 FVALERI 1049089  0 Jun 13 14:52 interpreter
```

FINISHED ▶ ✕ 🔍 ⚙

Took a few seconds. Last updated by anonymous at June 25 2016, 1:18:48 AM. (outdated)



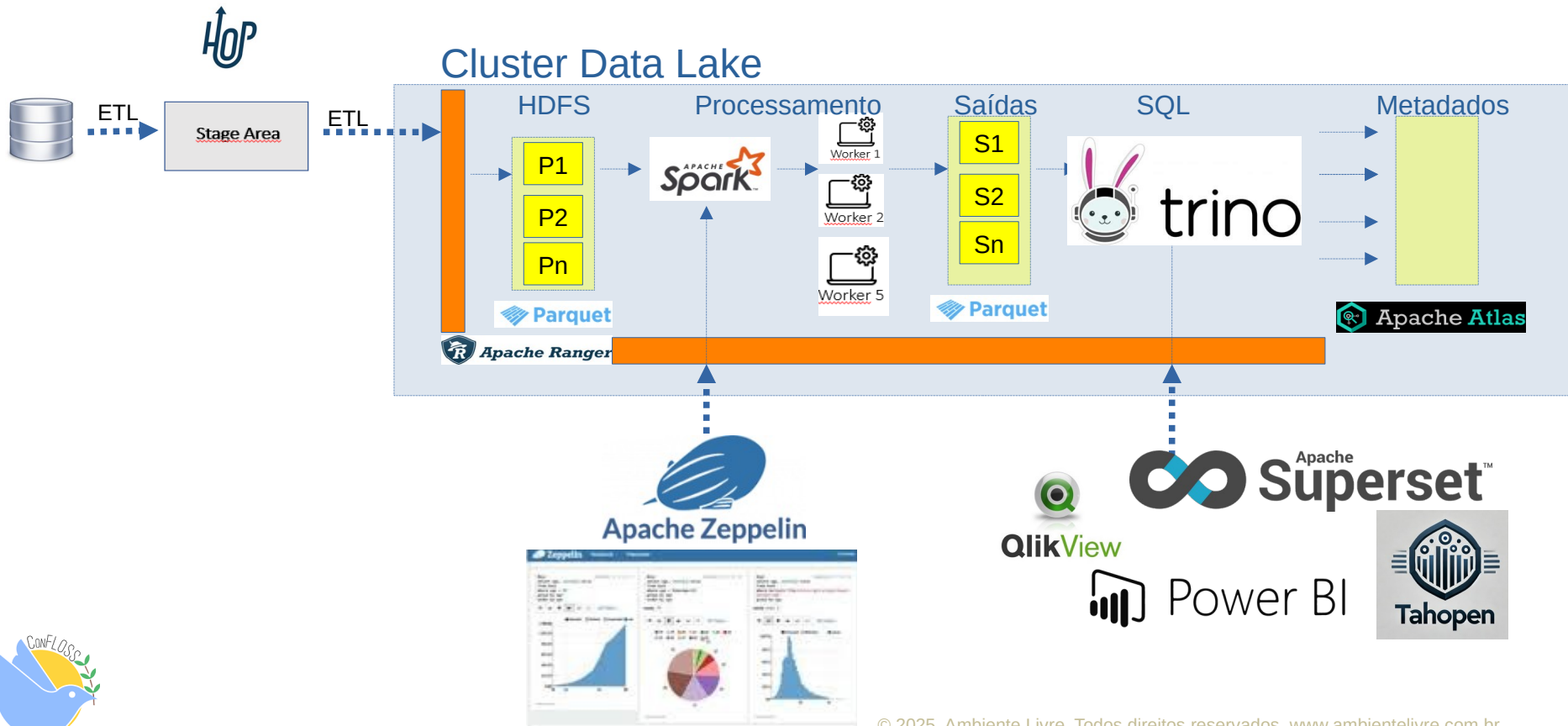
- O Apache Spark é suportado no grupo de intérpretes Zeppelin com Spark, que consiste em cinco intérpretes abaixo

Name	Class	Description
%spark	SparkInterpreter	Creates a SparkContext and provides a Scala environment
%spark.pyspark	PySparkInterpreter	Provides a Python environment
%spark.r	SparkRInterpreter	Provides an R environment with SparkR support
%spark.sql	SparkSQLInterpreter	Provides a SQL environment
%spark.dep	DepInterpreter	Dependency loader

- Na máquina Sandbox hortonworks temos o Interpreter **%spark** para spark 1.x e inferior e o interpreter **%spark2** para as versões superiores a 2.



Arquitetura Projeto Data Cloud



- **Documentação oficial Apache Zeppelin**
<http://zeppelin.apache.org/docs>

-



Obrigado

Marcio Junior Vieira

marcio@ambientelivre.com.br

@marviojvieira @ambientelivre

<https://www.linkedin.com/in/mvieira1/>

Blog: <http://blogs.ambientelivre.com.br/marcio/>

Slides:

<https://github.com/ambientelivre/labs>