

Ciência de Dados e Big Data

Recuperação da Informação na Web e em Redes Sociais

PUC-Minas IEC | Pós-Graduação Lato Sensu

Zilton Cordeiro Jr.



Projeto Final

Segunda metade da aula

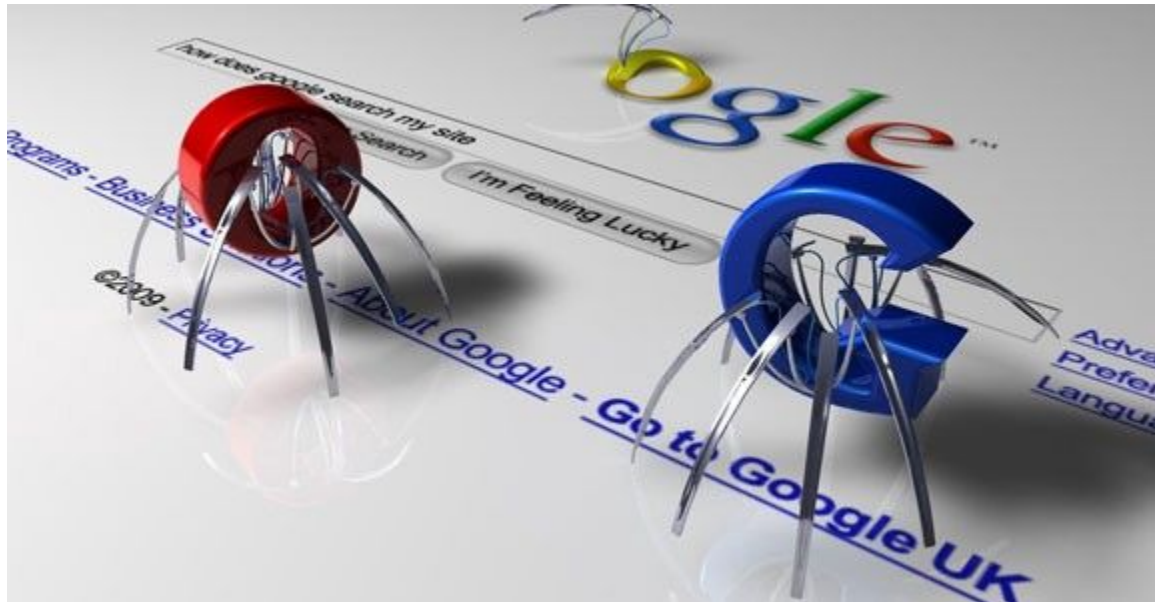


- ❖ O **projeto final** consiste em realizar um estudo da Web para um assunto real e de livre escolha.
 - Exemplos: Automóveis, moda, música, imóveis...
- ❖ Será necessário
 - Coletar dados em texto de redes sociais e sites da Web
 - Analisar o conteúdo textual obtido
 - Analisar dados de relacionamentos entre usuários (i.e. nas redes)
 - **Relatório final**
- ❖ **Data de Entrega**
 - 15° dia após a última aula às 23:59hrs

Busca Textual e Similaridade

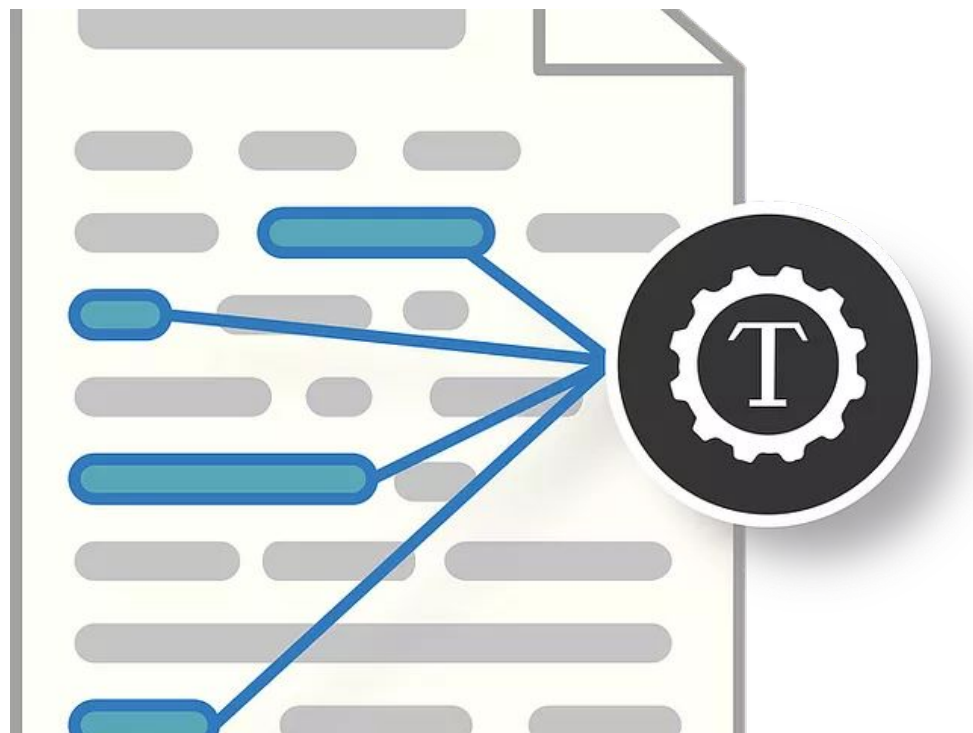


Coleta de Dados

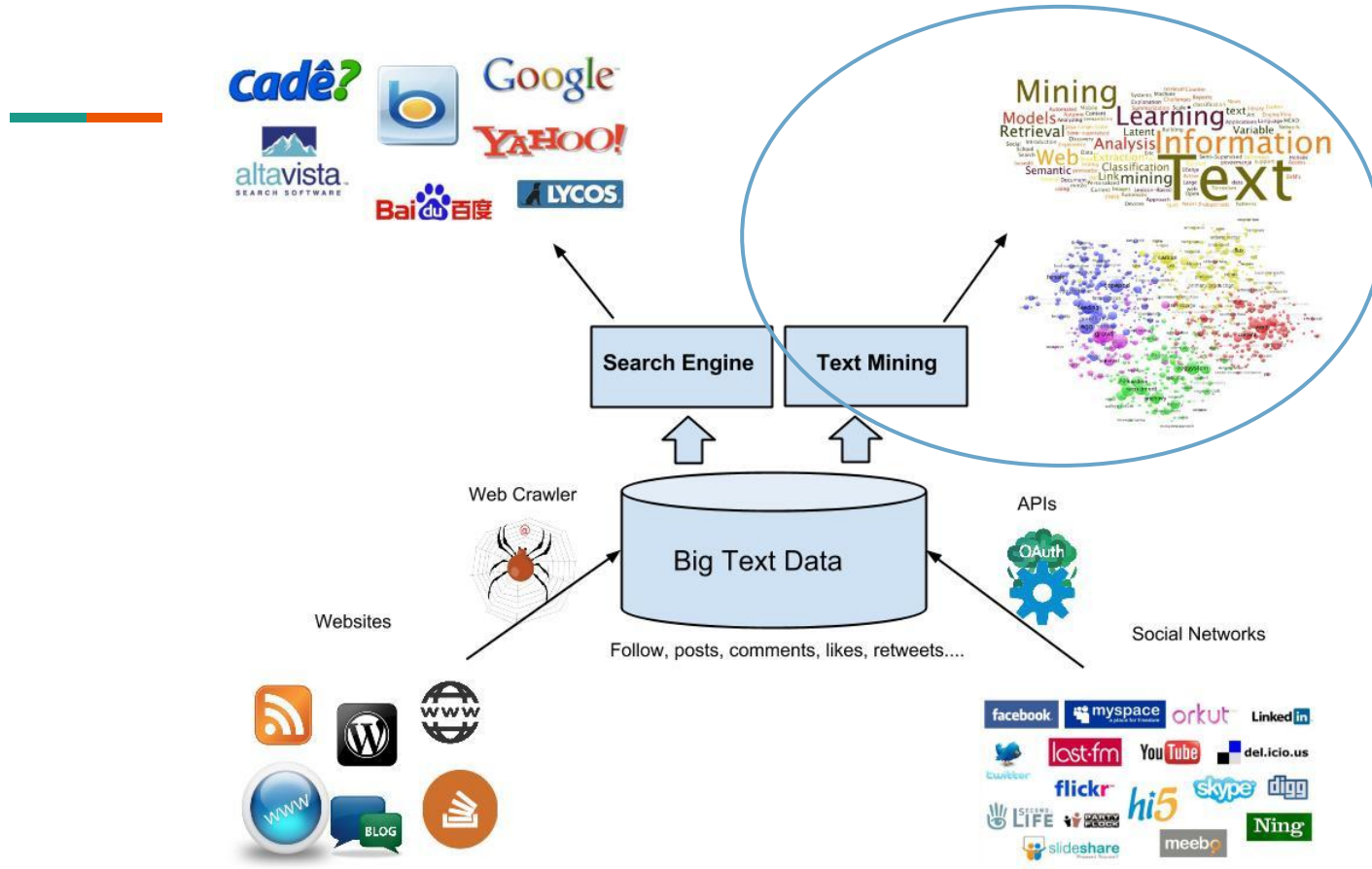




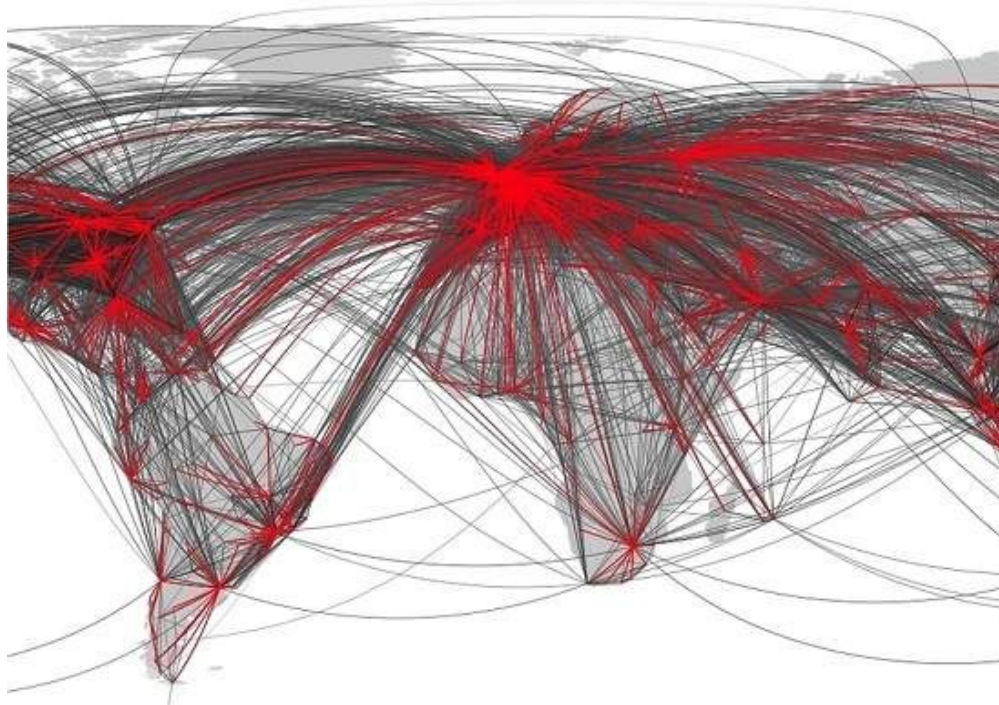
Mineração de Textos



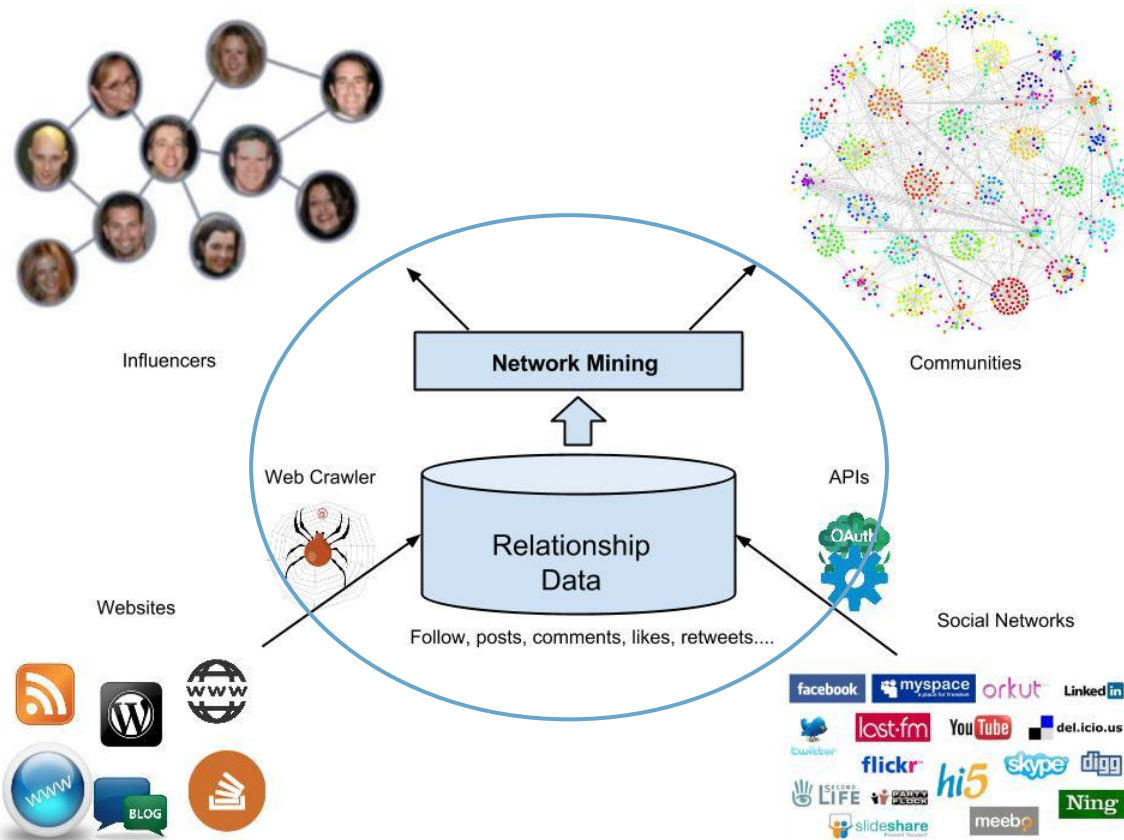
Mineração da Web e Redes Sociais




Grafos - Redes Complexas



Network Mining





Indexação, Busca e Mineração em plataformas de Big Data

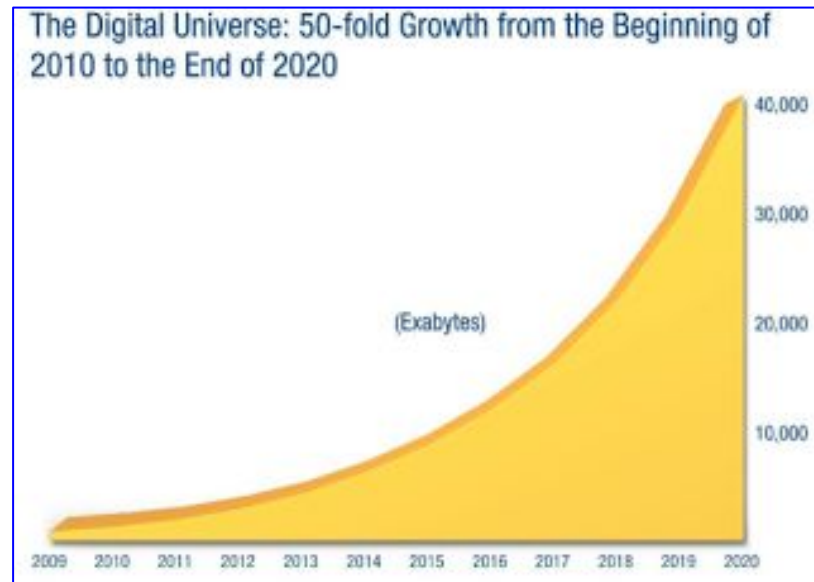
Big Data

❖ Volume de Dados

- Segundo a [Gartner](#)*: 2,2 milhões de terabytes de novos dados são criados todos os dias;
- A previsão é que até 2020 haja um total de 40 trilhões de gigabytes de dados no mundo.

* **Gartner** é uma empresa de consultoria fundada em 1979 por Gideon Gartner.

* A Gartner desenvolve tecnologias relacionadas a introspecção necessária para seus clientes tomarem suas decisões todos os dias.



Big Data

- ❖ **Pode ser entendido como:**
 - a captura;
 - o gerenciamento; e
 - a análise de dados.

- ❖ **Vai além de dados estruturados típicos**, que podem ser consultados por sistemas de gerenciamento de banco de dados relacional.
 - **Frequentemente são arquivos não estruturados:**
 - Vídeo digital;
 - Imagens;
 - Dados de sensores;
 - Arquivos de log; e
 - Qualquer dado não contido nos registros com campos pesquisáveis distintos.



Big Data

- ❖ Doug Laney deu uma definição para o Big Data com os três “V”:

- **V**olume,
- **V**elocidade e
- **V**ariiedade.



Big Data

❖ Volume

- Existem muitos fatores que contribuem para o aumento do volume de dados armazenados e trafegados:
 - Dados de transações;
 - Armazenados ao longo de vários anos;
 - Dados de texto
 - Áudio ou vídeo disponíveis em streaming nas mídias sociais; e
 - A crescente quantidade de dados coletados por sensores.



❑ No passado o volume de dados excessivo criou um problema de armazenamento.

- ❑ Mas, com os atuais custos de armazenamento decrescentes, outras questões surgem, incluindo:
 - ❑ Como determinar a relevância entre grandes volumes de dados?
 - ❑ Como criar valor a partir dessa relevância?

❖ Velocidade



- Significa o quão rápido os dados estão sendo produzidos e o quão rápido os dados devem ser tratados para atender às demandas.
- Reagir rápido o suficiente para lidar com a velocidade é um desafio para a maioria das organizações.



Big Data

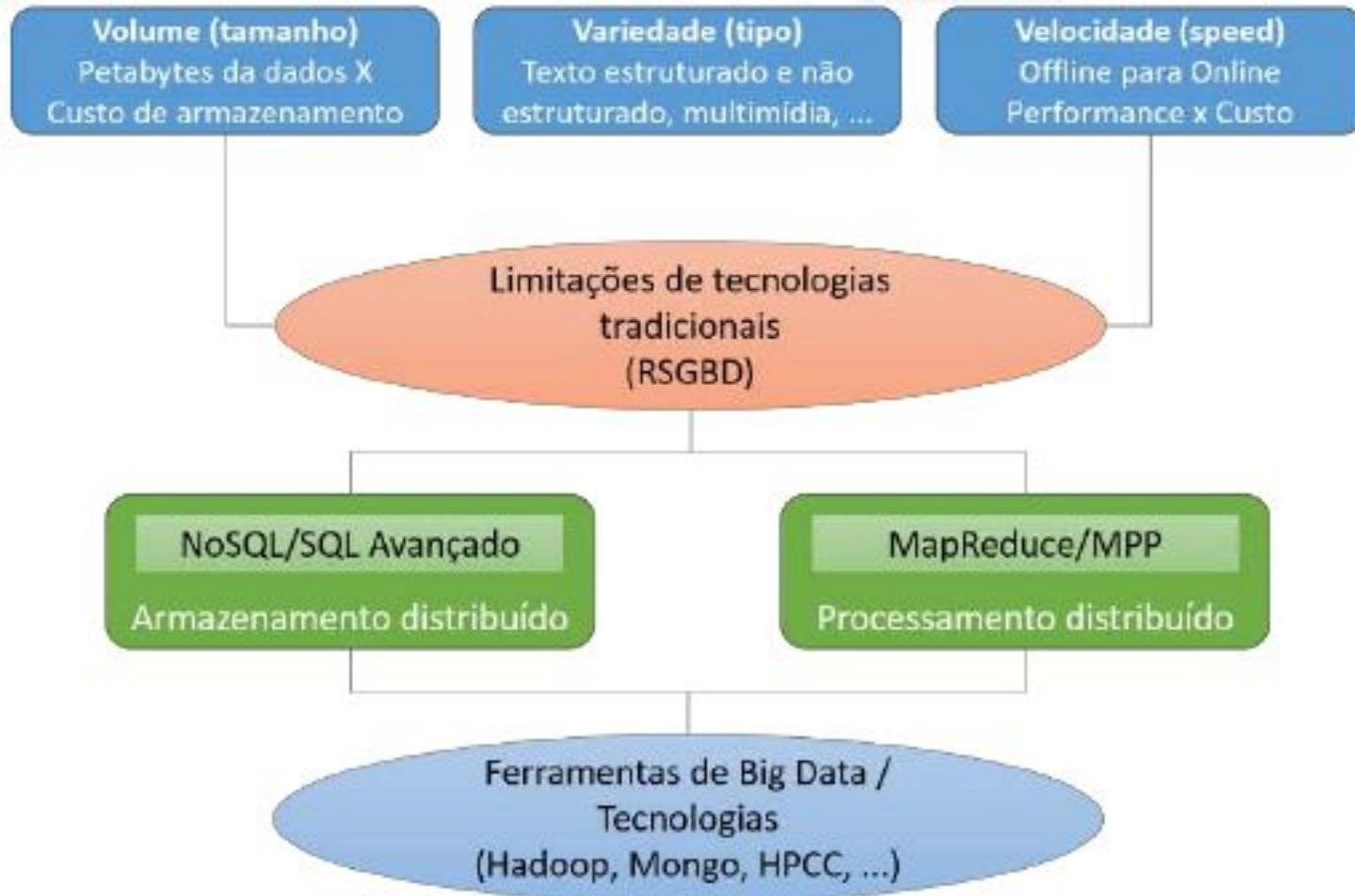
Variedade

- **Os dados de hoje vêm em todos os tipos de formatos:**
 - Bancos de dados tradicionais;
 - Arquivos de texto;
 - E-mail;
 - Medidores ou sensores de coleta de dados;
 - Vídeo;
 - Áudio;
 - Dados de ações do mercado e transações financeiras.

- ❑ Em algumas estimativas, 80% dos dados de uma organização não são numéricos! Mas, estes dados também precisam ser incluídos nas análises e nas tomadas de decisões das empresas.



Big Data

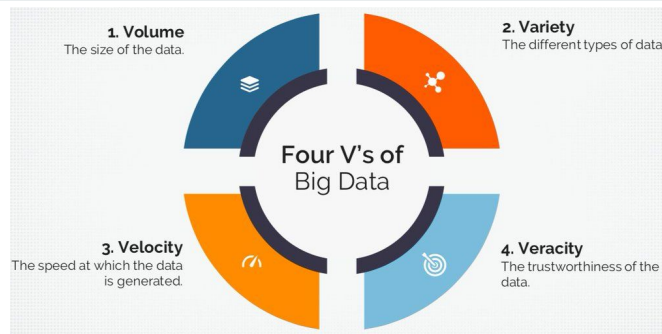


Big Data



❖ Veracidade

- Termo cunhado pela IBM, considerado o quarto “V”:
 - Representa a falta de confiabilidade inerente em algumas fontes de dados:
 - **Medir os sentimentos** dos clientes em mídias sociais é incerto por natureza, já que implicam uso do juízo humano.
 - No entanto, eles contêm **valiosas informações**.
- ❑ A necessidade de lidar com **dados imprecisos e incertos** é outra faceta de Big Data.
- ❑ Geralmente **resolvida** usando ferramentas e análises desenvolvidas para gerenciamento e **mineração de dados imprecisos**.
- ❑ É necessário **avaliar** as **inconsistências, incompletudes, ambiguidades**, latência e possíveis modelos de aproximação utilizados.
- ❑ Os **dados** podem ainda **perder** a **vigência**.
- ❑ **Verificar** se os **dados** são **consistentes** é extremamente necessário para qualquer análise de dados.



Big Data

❖ Visibilidade

- É a **relevância dos dados**. A organização está ciente de todos os dados que são gerados?
- Todos os **dados** gerados estão **disponíveis**?
- Os **dados** são, de fato, armazenados e ficam **visíveis para os analistas de dados**.



Big Data

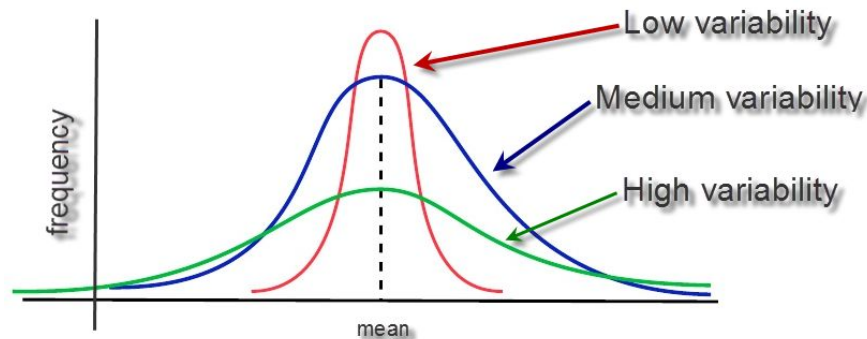
❖ Valor

- A Oracle introduziu **valor** como um atributo na definição de Big Data.
 - Big Data é, muitas vezes, caracterizado por uma "**densidade de valor relativamente baixa**".
 - Os dados recebidos na forma original, geralmente tem um **valor baixo em relação ao seu volume**.
 - Entretanto, um **valor elevado** pode ser **obtido pela análise de grandes volumes** destes mesmos dados.
 - **As informações geradas devem produzir algum valor para as organizações.**



❖ Variabilidade (e complexidade)

- A [SAS](#) apresentou variabilidade (e complexidade) como duas dimensões adicionais para Big Data.
 - **Variabilidade refere-se à variação nas taxas de fluxo de dados.**
 - Muitas vezes, a **velocidade de Big Data não é consistente e tem picos e depressões periódicas.**
 - **Complexidade refere-se ao fato de Big Data gerar ou receber informações através de uma multiplicidade de fontes.**
 - Isso impõe um desafio crucial: a **necessidade de se conectar, integrar, limpar e transformar os dados recebidos de diferentes fontes.**



Big Data

❖ Governança

- Ao decidir implementar ou não uma plataforma de big data, uma organização pode estar olhando novas fontes e novos tipos de elementos de dados nos quais a propriedade não está definida de forma clara.
 - Por exemplo:
 - No caso de assistência médica, é legal acessar dados de paciente para obter insight?
 - É correto mapear as despesas do cartão de crédito do cliente para sugerir novas compras?
- ❑ Regras semelhantes regem todos os segmentos de mercado.
- ❑ **Pode ser necessário redefinir ou modificar os processos de negócios de uma organização para que ela possa adquirir, armazenar e acessar dados externos.**



Big Data

❖ Pessoas

- É necessário ter **pessoas com aptidões específicas para entender, analisar os requisitos e manter uma solução de Big Data.**
- Envolve conhecimento do segmento de mercado;
- Domínio técnico sobre as ferramentas de Big Data; e
- Conhecimentos específicos de modelagem, estatística e outros.



Indexação de Dados



Indexing creates the “searchable” information that users will later use to find documents.

❖ Apache Lucene

- O Lucene contém apenas o núcleo do "motor" de busca.
 - O usuário do Lucene deve adicionar estas funcionalidades.
-
- ❑ Para o Lucene não importa a origem dos dados, seu formato ou mesmo a linguagem em que foi escrito, desde que esses dados possam ser convertido para texto.
 - ❑ Isto significa que o Lucene pode ser utilizado para indexar e buscar dados gravados em:
 - ❑ Arquivos;
 - ❑ Páginas web em servidores remotos;
 - ❑ Documentos gravados no sistema de arquivos local;
 - ❑ Arquivos textos;
 - ❑ Documentos Microsoft Word;
 - ❑ Documentos HTML ou arquivos PDF; ou
 - ❑ Qualquer outro formato do qual possa ser extraído informação textual.

❖ Solr: Ferramenta indexação e busca textual

- Criado em 2004 por Yonik Seeley como sistema de buscas do website da companhia [CNET Networks](#)
- Livre (open source). Doador para a Apache em 2006
- Baixa curva de aprendizado
- Altamente escalável
- REST-like API, configuração através de XML, **sem necessidade de codificação**



Busca em Big Data

❖ Quem usa o Solr



AT&T
Ticketmaster
Chegg
eBay
Magento
Comcast

Other Notable Users
Instagram
Netflix
Disney
Internet Archive
IBM Websphere Commerce
MTV Networks

Buy.com
The Echo Nest
Adobe
SAP Hybris
Bloomberg
Travelocity

Busca em Big Data

❖ Lidando com os dados



- “Schemaless” Fácil para começar a utilizar (possível criar esquemas próprios)
- Campos adicionados dinamicamente
- Componentes de text mining como bag of words, stemming...
- **Arquivos externos**, como **listas** de sinônimos, **stop words**, e palavras protegidas

Busca em Big Data

❖ Interface de administração: **localhost:8983**

The screenshot displays the Solr Admin UI. The browser address bar shows `127.0.0.1:8983/solr/#/`. The Solr logo is in the top left. The sidebar on the left contains the following links: Dashboard, Logging, Cloud, Core Admin, Java Properties, and Thread Dump. The 'Core Selector' dropdown menu is highlighted with a red box. The main content area is divided into three sections: Instance, Versions, and System. The Instance section shows the start time as '2 minutes ago'. The Versions section lists the following components and versions:

Component	Version
solr-spec	5.2.1
solr-impl	5.2.1 1684708 - shalin - 2015-06-10 23:20:13
lucene-spec	5.2.1
lucene-impl	5.2.1 1684708 - shalin - 2015-06-10 23:11:06

The System section on the right shows the following metrics:

- Physical Memory: 97.5%
- Swap Space: 12.5%
- File Descriptor Count: 5.2%

A red text annotation with an arrow points to the 'Core Selector' dropdown menu, stating: 'Selecione um índice salvo para mais opções'.

Solr - Busca (Query)

Termos de busca →

Filtragem →

Ordenação →

Paginação →

Definição de campos →

Parâmetros além dos disponíveis na interface →

Request-Handler (qt)

/select

— common —

q

late

fq

id asc

start, rows

2 10

fl

Document_body_text Title Category

df

Raw Query Parameters

key1=val1&key2=val2

http://127.0.0.1:8983/solr/docs_knime_shard1_replica1

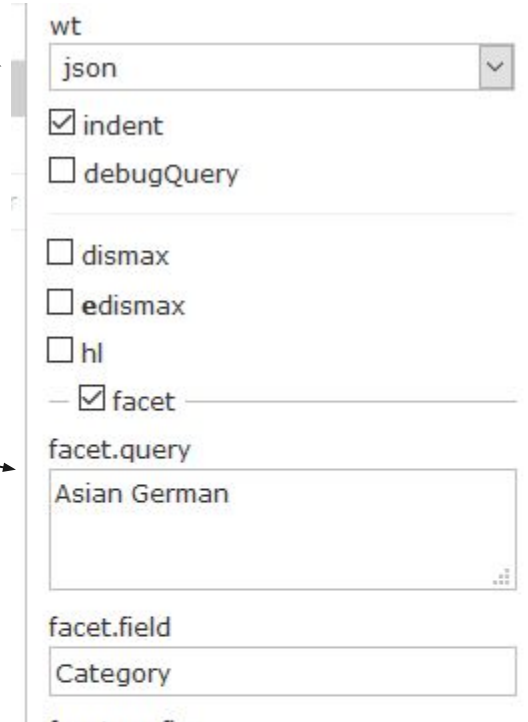
```
{
  "responseHeader": {
    "status": 0,
    "QTime": 3,
    "params": {
      "facet": "true",
      "fl": "Document_body_text Title Category",
      "sort": "id asc",
      "indent": "true",
      "facet.query": "(Asian German Fast) NOT Food",
      "start": "2",
      "q": "late",
      "_": "1492977041001",
      "facet.field": "Category",
      "wt": "json",
      "rows": "10"
    }
  },
  "response": {
    "numFound": 6,
    "start": 2,
    "docs": [
      {
```


Solr - Busca (Query)

Formato de retorno

Buscas facetadas
(agregações)

Definir campo



The image shows a screenshot of the Solr Admin UI configuration for a query. It includes several input fields and checkboxes. The 'wt' (write method) is set to 'json'. The 'indent' checkbox is checked. The 'facet.query' field contains 'Asian German'. The 'facet.field' field contains 'Category'. The 'facet' checkbox is checked.

wt
json
☒ indent
☐ debugQuery
☐ dismax
☐ edismax
☐ hl
— ☒ facet —
facet.query
Asian German
facet.field
Category

```
},  
"facet_counts": {  
  "facet_queries": {  
    "Asian German": 3  
  },  
  "facet_fields": {  
    "Category": [  
      "Fast Food",  
      3,  
      "German Cuisine",  
      2,  
      "Asian",  
      1  
    ]  
  },  
}
```

Solr - Busca

http://127.0.0.1:8983/solr/docs_knime_shard1_replica1/select?q=late&sort=id+asc&start=2&rows=10&fl=Document_body_text+Tit

**Consulta via
URL**

← ⓘ 127.0.0.1:8983/solr/docs_knime_shard1_replica1/select?q=late&

```
{
  "responseHeader":{
    "status":0,
    "QTime":2,
    "params":{
      "facet":"true",
      "fl":"Document_body_text Title Category",
      "sort":"id asc",
      "indent":"true",
      "facet.query":"Asian German",
      "start":"2",
      "q":"late",
      "facet.field":"Category",
      "wt":"json",
      "rows":"10"}},
  "response":{"numFound":6,"start":2,"docs":[
    {
      "Title":["Solr - Busca"]
```

Sorl - Indexação

❖ Criação de índice

```
bin/solr create -c meu_novo_indice -d minha_nova_config -s 1 -rf 1 -p 8983
```

Executar

```
/opt/lucidworks-hdpsearch/solr/bin/solr create -c meu_novo_indice -d minha_nova_config -s 1 -rf 1 -p 8983
```

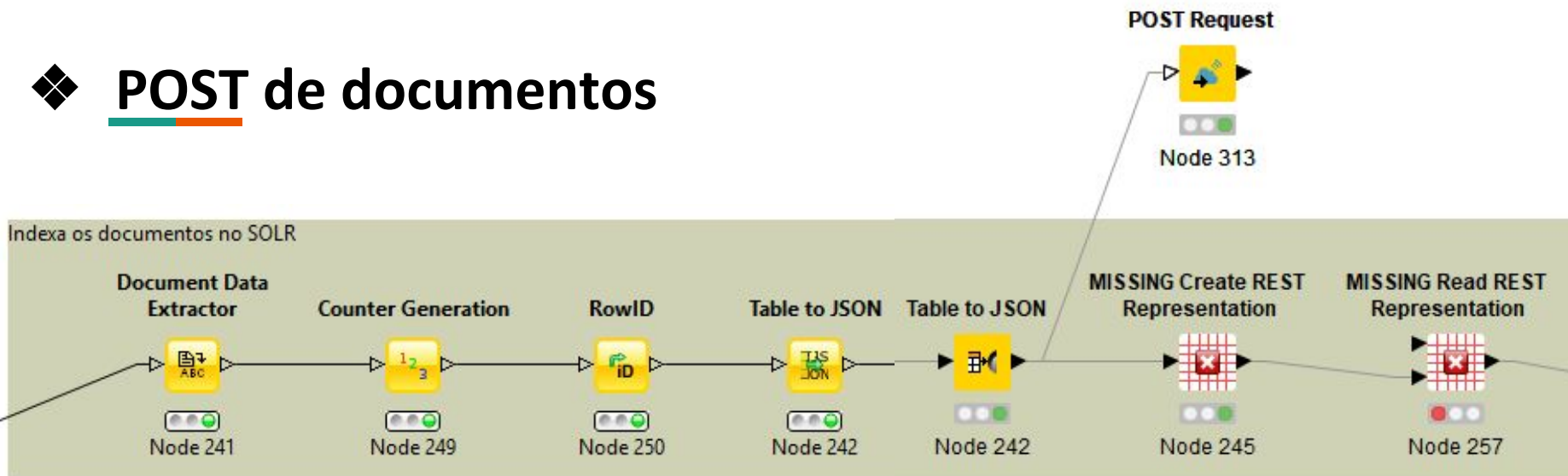
Solr - Indexação

❖ Requisição realizada..

```
Creating new collection 'meu_novo_indice' using command:  
http://10.0.2.15:8983/solr/admin/collections?action=CREATE&n
```

Solr - Indexação

❖ POST de documentos



Em qualquer linguagem, basta enviar os documentos em json via método POST para:

http://127.0.0.1:8983/solr/NOME_DO_SEU_INDICE/update/json/docs?commit=true

Sorl - Indexação

Dialog - 0:313 - POST Request

File

Request Body	Response Headers	Flow Variables	Job Manager Selection	Memory Policy
Connection Settings		Authentication		Request Headers
<input checked="" type="radio"/> URL:	http://127.0.0.1:8983/solr/tweets_2_shard1_replica1/update/json/docs?commit=true			
<input type="radio"/> URL column:	{JS ON add			
<input type="checkbox"/> Delay (ms):	0			
Concurrency:	1			

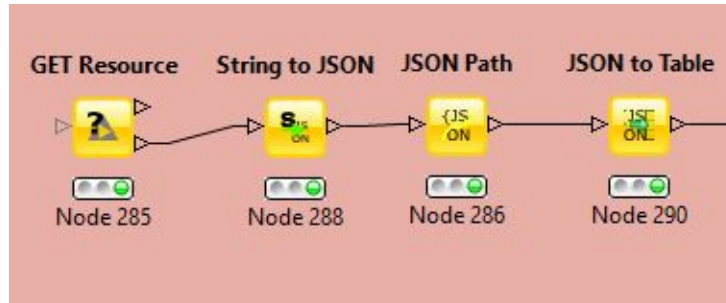
Dialog - 0:313 - POST Request

File

Connection Settings	Response Headers	Flow
Request Body		
<input checked="" type="radio"/> Use column's content as body {JS ON add		
<input type="radio"/> Use constant body		

Solr - Busca

❖ GET - Busca (similaridade)

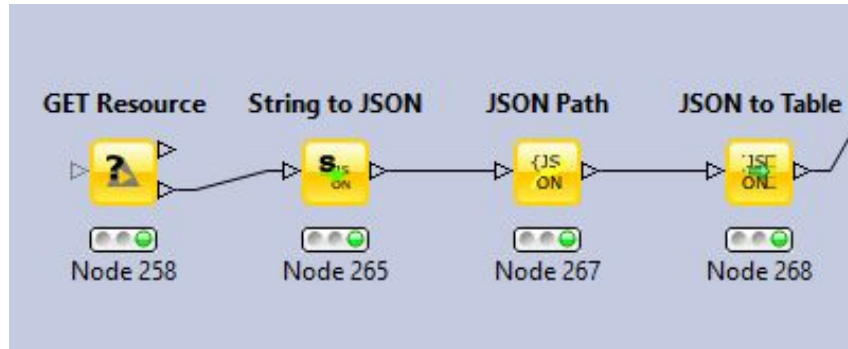


Em qualquer linguagem, basta acessar a URL e parâmetros (mesma URL que a interface gera):

http://127.0.0.1:8983/solr/NOME_DO_INDICE/select?q=%22great+food%22&wt=json&indent=true

Solr - Busca

❖ GET - Contagem de palavras



Em qualquer linguagem, basta acessar a URL e parâmetros (mesma URL que a interface gera):

http://127.0.0.1:8983/solr/NOME_DO_INDICE/select?q=%3A*&wt=json&rows=0&facet=true&facet.field=_text_&facet.limit=80

❖ Elasticsearch

- Servidor de buscas distribuído baseado no Apache Lucene.
- Disponibilizado sobre os termos Apache License.
- Desenvolvido em Java e possui código aberto liberado sob os termos da Licença Apache.



Busca em Big Data

❖ O **Elasticsearch** realiza buscas por Índice Invertido:

- No momento em que um documento é indexado, o Elasticsearch separa todos os seus termos em Tokens.
 - Em seguida ele faz uma medição para definir quais tokens são relevantes, eliminando assim artigos, preposições, etc.
 - O próximo passo do Elasticsearch é organizar os tokens em um índice e informar em cada token quais documentos contém esse token.
 - Quando uma busca for feita ela agirá sobre esse índice invertido ao invés de vasculhar cada documento individualmente, procurando pelos termos buscados.
- ☐ Esse processo de indexação é o que torna o Elasticsearch um motor de busca em semi-tempo-real.

❖ ElasticSearch



- O Elasticsearch suporta um grande volume de dados sem perder performance.
- Pode ser implementado em qualquer sistema independentemente da plataforma, por fornecer uma API REST.
- Ferramenta é altamente escalável, podendo ir de um servidor a muitos servidores simultâneos.



Busca em Big Data

❖ ElasticSearch: Utiliza?



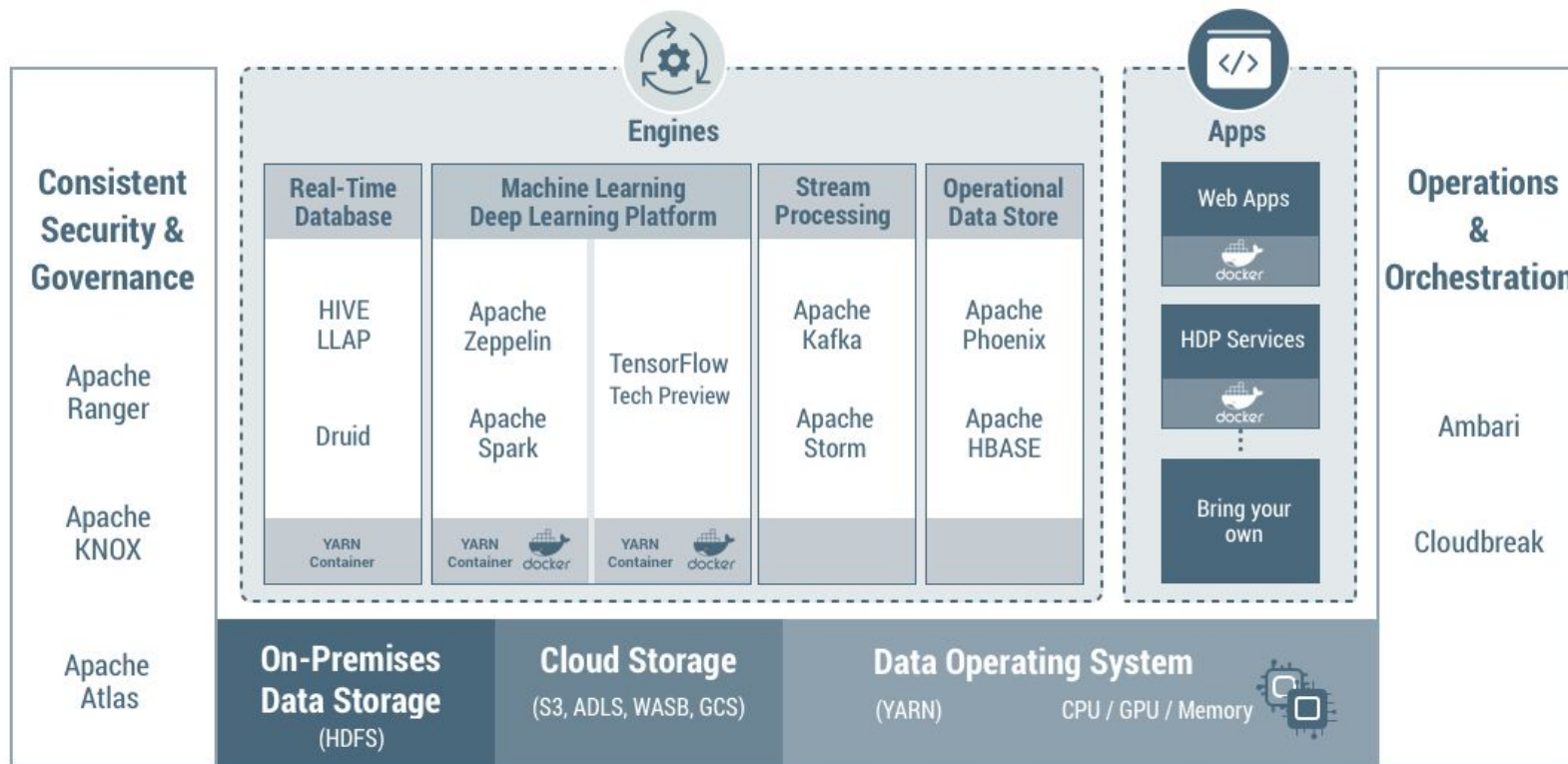
❖ Hortonworks Data Platform (HDP)

- É uma estrutura de código aberto para armazenamento e processamento distribuídos de grandes conjuntos de dados de várias fontes.



Big Data

❖ Hortonworks Data Platform (HDP)

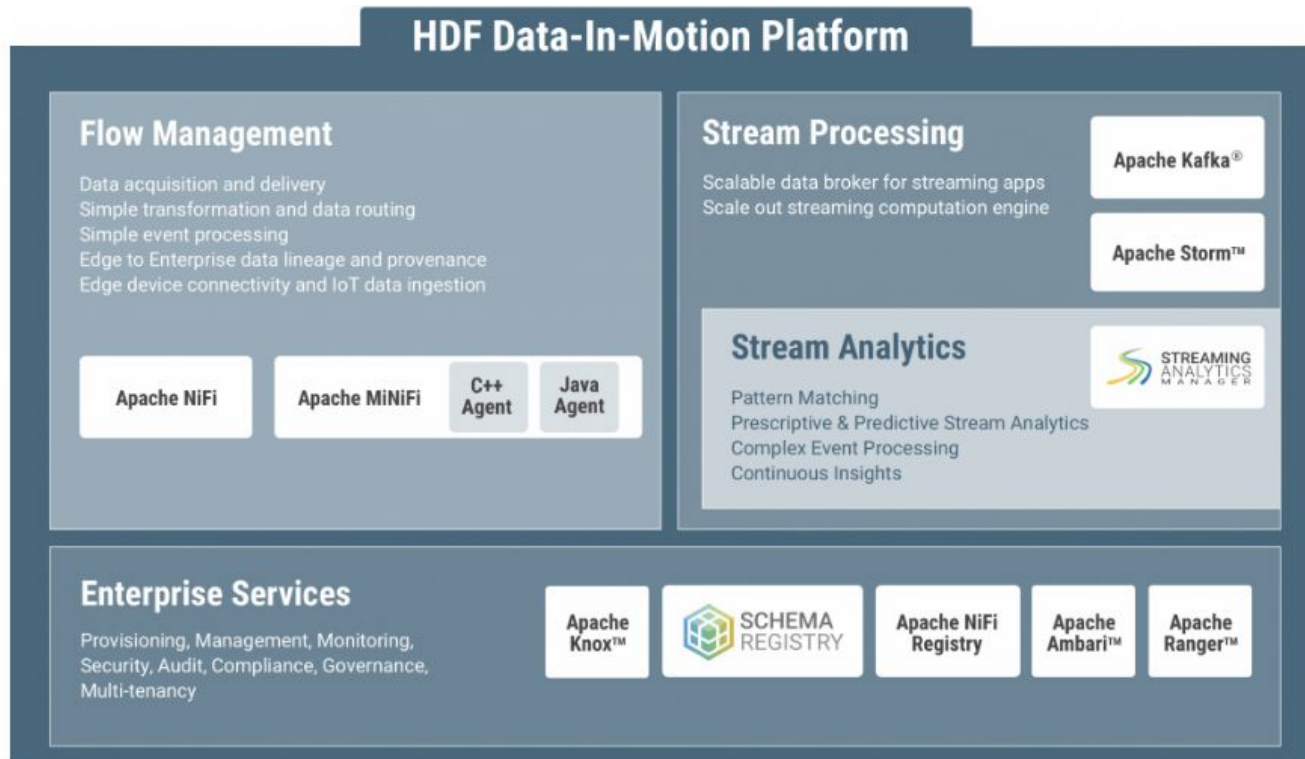


❖ Hortonworks DataFlow (HDF)

- É uma plataforma analítica de streaming escalável e em tempo real que ingere, organiza e analisa dados para obter informações importantes e inteligência prática imediata.



❖ Hortonworks DataFlow (HDF)



❖ Apache Spark

- É um sistema de processamento distribuído de código aberto usado normalmente para cargas de trabalho de big data.
- O Apache Spark utiliza o armazenamento em cache na memória e a execução otimizada para obter alta performance.
- Oferece suporte a:
 - Processamento geral de lotes;
 - Análise de streaming;
 - Machine Learning
 - Bancos de dados gráficos; e
 - Consultas ad hoc.





Toolkit para Ciência de Dados

Toolkit para Ciência de Dados

❖ Anaconda ([download](#))

- Distribuição de alta performance para Python, R e Scala
- Inclui mais de 100 bibliotecas e recursos necessários para projetos de Data Science e Machine Learning
- **Jupyter Notebook**, a IDE Spyder, NumPy, Pandas, Scikit-learn...



Toolkit para Ciência de Dados

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

Home

Environments

Learning

Community


Documentation


Developer Blog


Feedback


Twitter YouTube GitHub


Applications on root Channels Refresh



jupyter
notebook
4.3.1
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch



IPyT
qtconsole
4.2.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch


spyder
3.1.2
Scientific Python Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
Launch


anaconda-fusion
1.0.2
Integration between Excel ® and Anaconda via Notebooks. Run data science Functions, interact with results and create advanced visualizations in a code-free app inside Excel
Install


glueviz
0.9.1
Multidimensional data visualization across files. Explore relationships within and among related datasets.
Launch


orange3
3.4.1
Launch

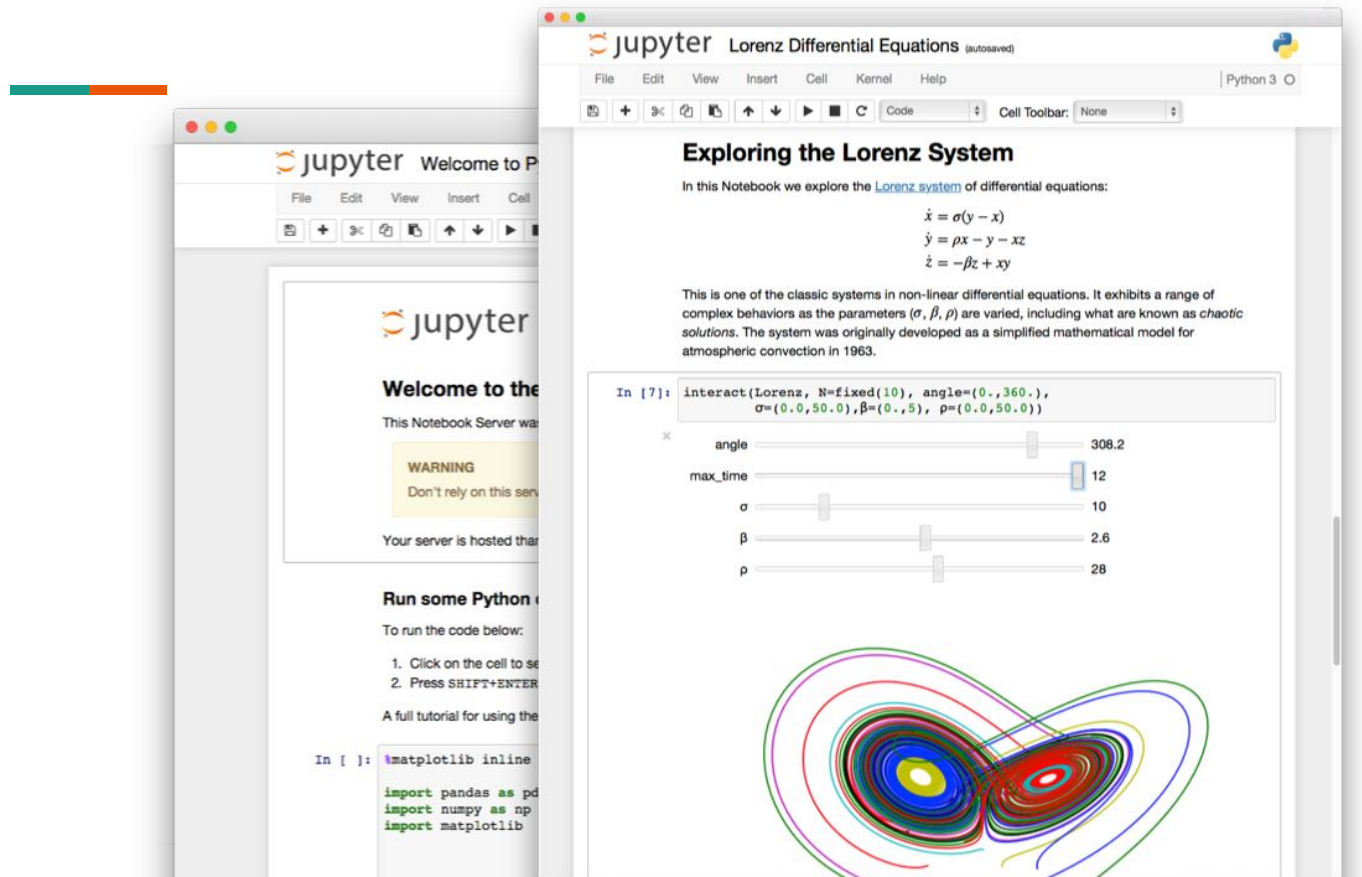

rstudio
1.0.136
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.
Launch

❖ Jupyter Notebook (Já vem com o Anaconda)

- É uma aplicação (open source) que permite criar e compartilhar documentos com código dinâmico, visualizações e textos explicativos.
- Suporte a elementos HTML e executa no browser - bem melhor que em linhas de comando :-)



Toolkit para Ciência de Dados



The image displays three overlapping Jupyter Notebook windows. The background window shows the 'Welcome to Jupyter' page with a warning about server reliability and instructions on how to run Python code. The middle window shows the 'Exploring the Lorenz System' notebook, which includes the Lorenz equations and a plot of the Lorenz attractor. The foreground window shows the same notebook with interactive sliders for parameters like angle, max_time, sigma, beta, and rho, and a corresponding plot of the Lorenz attractor.

Exploring the Lorenz System

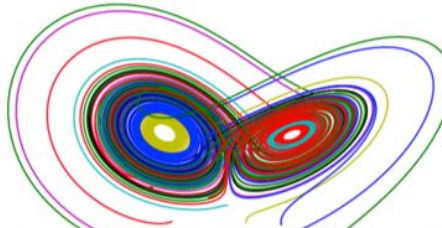
In this Notebook we explore the [Lorenz system](#) of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

In [7]: `interact(Lorenz, N=fixed(10), angle=(0.,360.),
sigma=(0.0,50.0),beta=(0.,5), rho=(0.0,50.0))`

angle 308.2
max_time 12
sigma 10
beta 2.6
rho 28



The plot shows the Lorenz attractor, a complex, chaotic system trajectory in the x-y-z phase space, rendered with multiple colored lines (red, blue, green, yellow, cyan) to illustrate the system's behavior over time.

❖ Para iniciar o jupyter

Executa

```
jupyter notebook --ip=0.0.0.0 --port=8889
```

```
--NotebookApp.token="
```

Toolkit para Ciência de Dados

❖ Para iniciar o jupyter

```
[root@sandbox jupyter]# ./start_jupyter.sh
[I 22:07:22.837 NotebookApp] Serving notebooks from local directory: /media/sf_storage
[I 22:07:22.837 NotebookApp] 0 active kernels
[I 22:07:22.837 NotebookApp] The Jupyter Notebook is running at: http://0.0.0.0:8889/
[I 22:07:22.837 NotebookApp] Use Control-C to stop this server and shut down all kernels
[W 22:07:22.869 NotebookApp] No web browser found: could not locate runnable browser.
[C 22:07:22.872 NotebookApp]
```

Copy/paste this URL into your browser when you connect for the first time,
to login with a token:

`http://0.0.0.0:8889/?token=e0a26df9852868d86915a049abf9153a2ca17ea46fa5f4dd`

Colar o link no navegador

No navegador substituir 0.0.0.0 por 127.0.0.1

Toolkit para Ciência de Dados

❖ Home

127.0.0.1:8889/tree?token=e0a26df9852868d86915a049abf9153a2ca17ea46fa5f4dd



Files

Running

Clusters

Select items to perform actions on them.



Crawling - FeedParser.ipynb



Crawling - Scrap with login.ipynb



Crawling - TextExtraction.ipynb



Querying.ipynb



start_jupyter.sh

Python + Solr: Busca

PySolr

Embora seja possível fazer buscas e interações via requisições REST (GET, POST), com a lib pysolr fica ainda mais fácil e conveniente!

```
In [3]: from __future__ import print_function
import pysolr, json
```

```
In [4]: solr = pysolr.Solr("http://localhost:8983/solr/tweets_pucminas/", timeout=10)
```

```
In [15]: results = solr.search(q="Tweet:ferrovia")
```

```
In [16]: print("{0} result(s)".format(len(results)))
```

1 result(s).

```
In [18]: for result in results:
print(result['Tweet'])
```

['No próximo dia 28/03 será leiloado o trecho central e sul da Ferrovia Norte-Sul. Este corredor que cortará o Brasil, cria uma nova espinha dorsal na logística de transporte de produção em nosso país, gerando uma série de benefícios econômicos e sociais a todos os brasileiros. <https://t.co/h92SsJr3oA>']

❖ Banana ([Download](#))

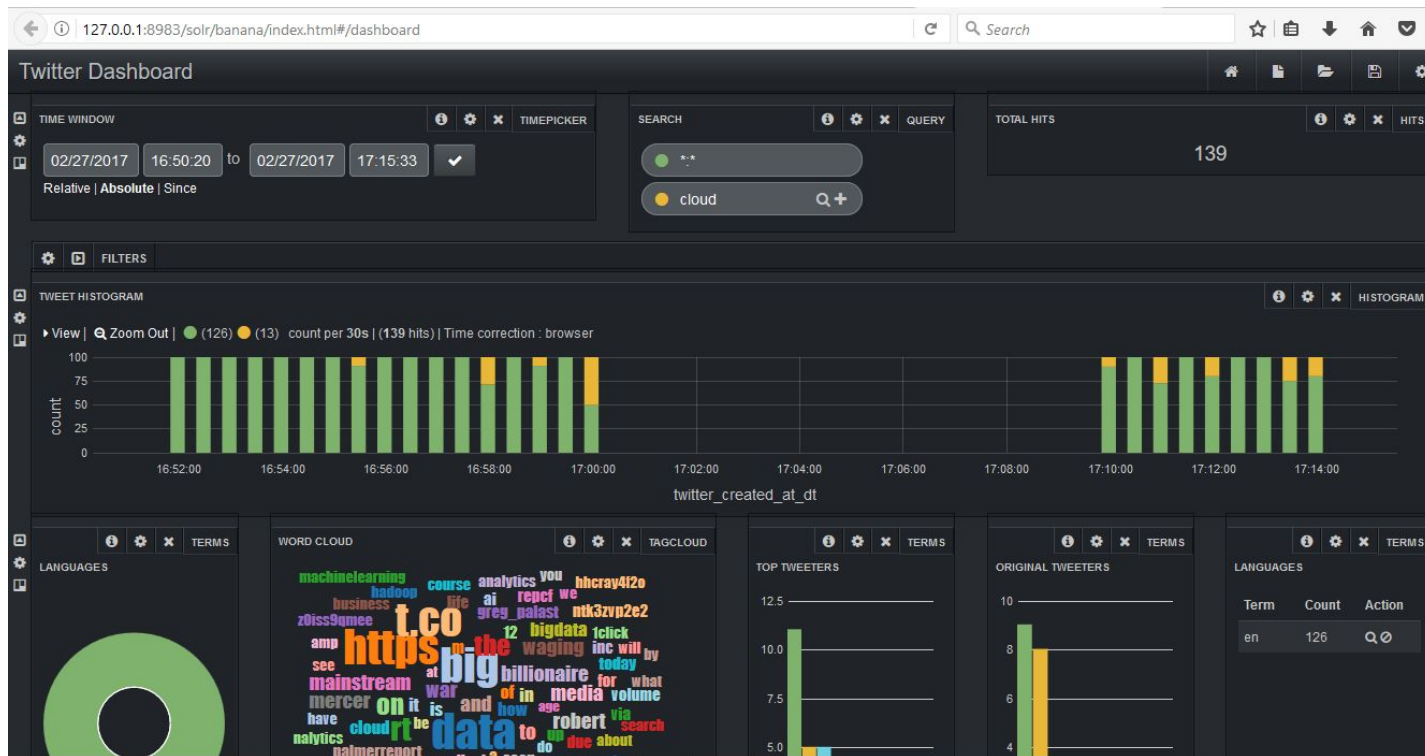
- Configuração de interfaces flexíveis que se conectam diretamente no Solr
- Inclui painéis que utilizam a poderosa biblioteca de visualizações em javascript **D3.js**
- Já vem com Solr-HDP

Banana

The logo for Lucidworks, featuring a red square icon with a white 'L' shape inside, followed by the word 'Lucidworks' in white text on a dark blue background.

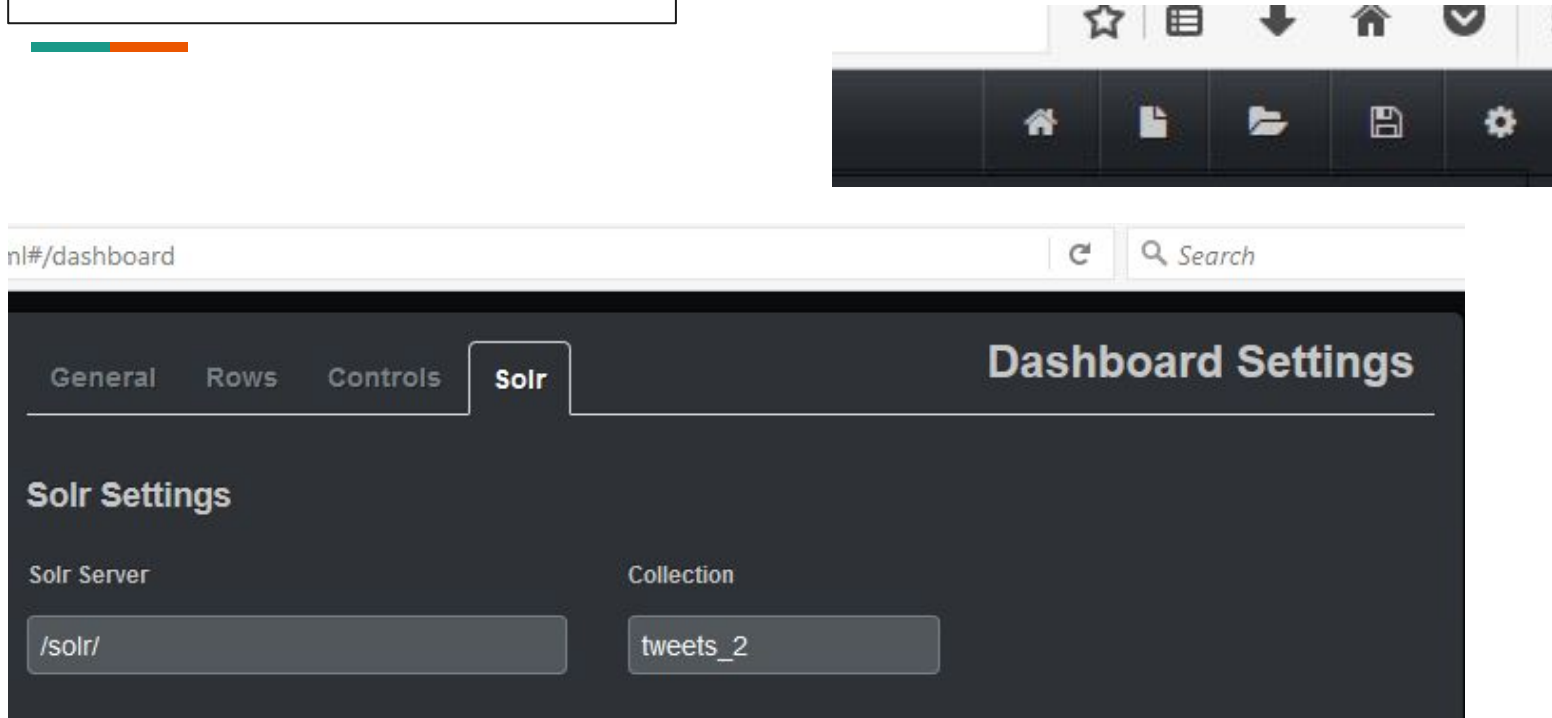
Sorl - Visualização

<http://127.0.0.1:8983/solr/banana/index.html>



Solr - Visualização

Configurações via interface



Sorl - Visualização

Configurações via interface

Adição de painéis:
Configurar linha -> Add Panel

The image shows two overlapping screenshots of the Sorl web interface. The background screenshot displays the 'tagcloud' panel configuration. At the top, a dropdown menu is set to 'tagcloud' with a note: 'Note: This row is full, new panels will wrap to a new line. You should add another row.' Below this is a section titled 'Experimental // Display the tag cloud of the top N words from a specified field.' It contains several settings: 'Title' (empty text field), 'Span' (dropdown set to 4), 'Editable' (checkbox checked), 'Inspect' (checkbox checked), 'Field' (empty text field), 'Number of tags' (dropdown set to 10), 'Text Alignment' (dropdown set to 'vertical and horizontal'), and 'Font Scale' (dropdown set to 1). A 'Queries' section has a checkbox 'Display query when Inspect' which is checked, and a 'Panel Query' field containing '*: *'. At the bottom right is a green 'Add Panel' button.

The foreground screenshot is a modal dialog titled 'Add Panel'. It has three tabs: 'General', 'Panels', and 'Add Panel', with the 'Add Panel' tab selected. Inside the dialog, there is a 'Select Panel Type' dropdown menu, which is currently empty. A note below the dropdown says: 'Note: This row is full, new panels will wrap to a new line. You should add another row.' At the bottom of the dialog, there is a green 'Add Panel' button and a red button partially visible.

Mineração via HIVE (distributed storage using SQL)

http://127.0.0.1:9995

Zeppelin - Oferece uma camada que interpreta várias sintaxes como SQL, Scala, Cassandra e entre outros, além disso te permite criar visualização de dados rapidamente através dos resultados obtidos pelos interpretadores

Zeppelin (~ jupyter)

Já vem instalado na HDP com interpretadores que permitem rodar scripts direto na infraestrutura como %pyspark, %hive, %sql ...



The screenshot shows the Zeppelin web interface in a browser. The address bar displays '127.0.0.1:9995/#/'. The page has a blue header with the Zeppelin logo and navigation links for 'Notebook' and 'Interpreter'. The main content area features a 'Welcome to Zeppelin!' message, a description of Zeppelin as a web-based notebook, and a list of available notebooks under the heading 'Notebook'. The list includes 'AON Demo', 'Australian Dataset (Hive example)', 'Australian Dataset (SparkSQL example)', 'Hello World Tutorial', 'IoT Data Analysis (Keynote Demo)', 'KNIME Docs - Sentiment', and 'KNIME Docs - Sentiment'.

127.0.0.1:9995/#/

Zeppelin Notebook Interpreter

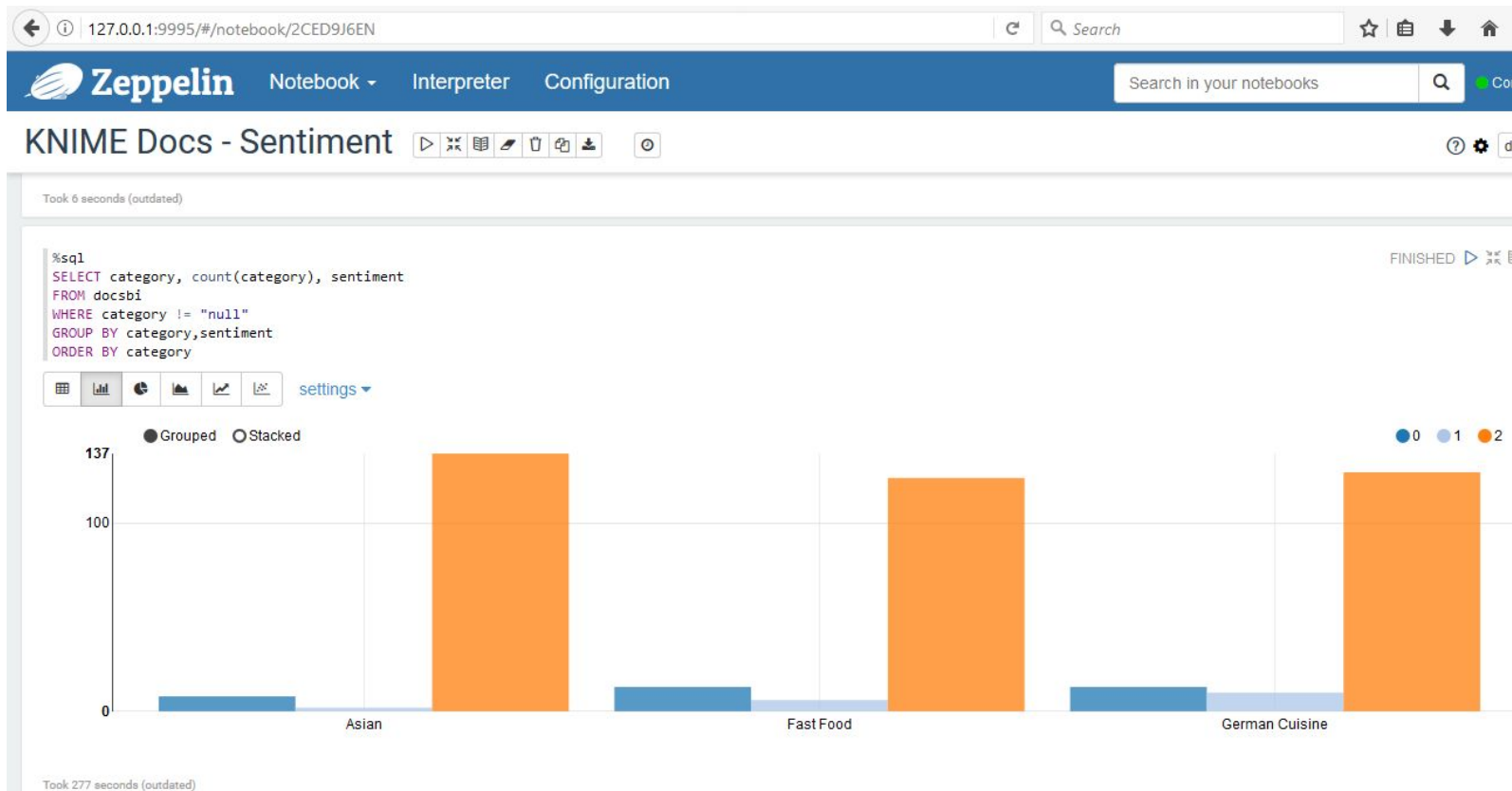
Welcome to Zeppelin!

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

Notebook

- Import note
- Create new note
 - AON Demo
 - Australian Dataset (Hive example)
 - Australian Dataset (SparkSQL example)
 - Hello World Tutorial
 - IoT Data Analysis (Keynote Demo)
 - KNIME Docs - Sentiment
 - KNIME Docs - Sentiment

Mineração via HIVE (distributed storage using SQL)





Tutorial on Twitter Sentiment Analysis and n-gram with Hadoop and Hive SQL

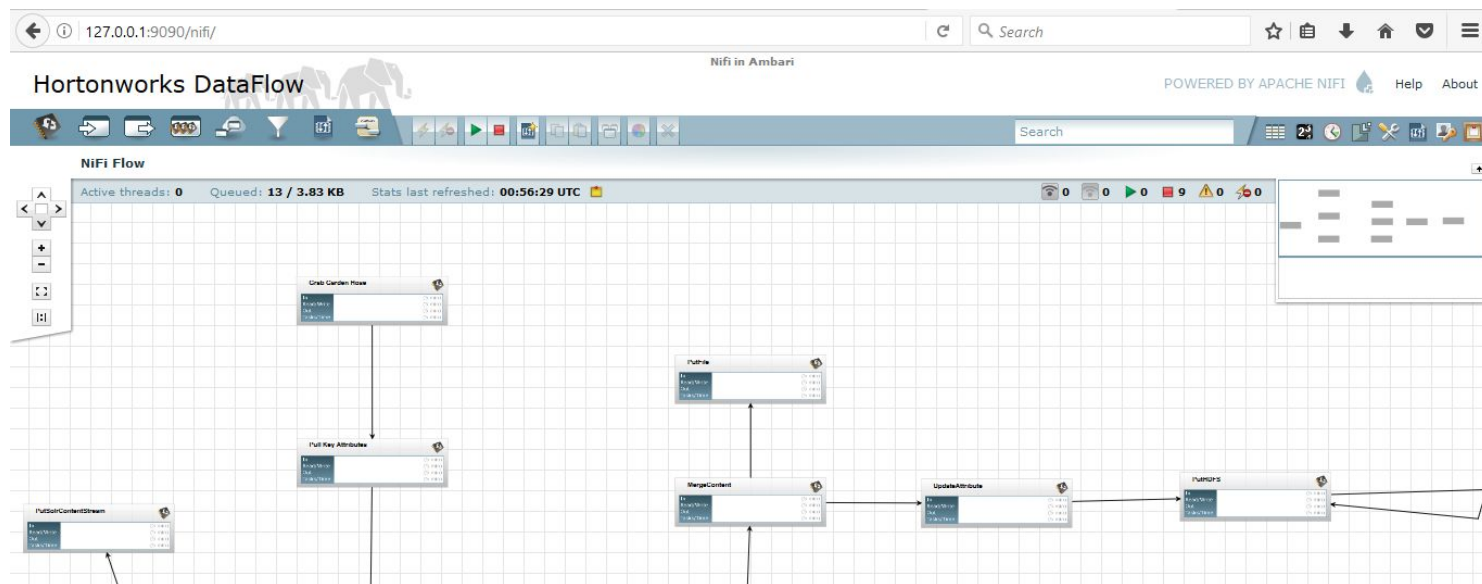
<https://gist.github.com/umbertogriffo/a512baaf63ce0797e175>

Sobre tabelas ORC

<https://br.hortonworks.com/hadoop-tutorial/using-hive-with-orc-from-apache-spark/>

NIFI - O “KNIME” da HDP

Inicie o servidor NIFI no Ambari e acesse:
<http://127.0.0.1:9090/nifi/>



Tutorial Hortonworks



Analyzing Social Media and Customer Sentiment with Apache NIFI and HDP Search

<https://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-sentiment-data/>

❖ Plano de Ensino

- **Unidade 01:** Conceitos de inteligência competitiva e coletiva, crowdsourcing e redes sociais. Recuperação da informação e Máquinas de busca. Desafios da Mineração na web e nas redes. Exemplos de Projetos da disciplina.
- **Unidade 02:** Algoritmos e soluções para problemas de busca e extração de informação da WWW. Ferramenta e prática de processamento textual e recuperação de informação.
- **Unidade 03:** Tipos de coleta, arquitetura e componentes de coletores Web. Ferramenta e prática de coleta de dados na Web.

❖ **Plano de Ensino**

- **Unidade 04:** Aprofundando na mineração de texto e linguagem natural. Algoritmos e soluções para a análise da informação presente nas redes sociais online e em sites de conteúdo. Ferramenta e prática de mineração de texto.
- **Unidade 05:** Caracterização de redes sociais: Tipologia, características e representações gráficas. Algoritmos estocásticos, análise de redes complexas. Ferramenta e prática de mineração de redes complexas.
- **Unidade 06:** Indexação, Busca e Mineração em plataforma de Big Data

❖ Teórico e Prático

O conteúdo estudado será exercitado em práticas utilizando ferramentas de mineração de texto e busca.

As aulas práticas serão avaliadas e em cada prática uma tarefa deverá ser realizada de maneira autônoma. **40 pontos.**

O Projeto Final será formado por conceitos discutidos e aplicados nas aulas, com adaptações individuais para um caso de uso real. O resultado das tarefas práticas poderá ser reaproveitado. **60 Pontos**

Projeto Final

- ❖ O **projeto final** consiste em realizar um estudo da Web para um assunto real e de livre escolha.
 - Exemplos: Automóveis, moda, música, imóveis...
- ❖ Será necessário
 - Coletar dados em texto de redes sociais e sites da Web
 - Analisar o conteúdo textual obtido
 - Analisar dados de relacionamentos entre usuários (i.e. nas redes)
 - Relatório final
- ❖ **Data de Entrega**
 - 15° dia após a última aula às 23:59hrs