Open Source Data Science

Elaborando uma plataforma de Big Data & Analytics 100% Open Source com apoio do Pentaho.

Palestrante: Marcio Junior Vieira
CEO e Data Scientist na Ambiente Livre
marcio@ambientelivre.com.br

Marcio Junior Vieira

- 17 anos de experiência em informática, vivência em desenvolvimento e análise de sistemas de Gestão empresarial e Analise de Dados.
- 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: CONISLI, SOLISC, FISL, LATINOWARE, SFD, JDBR, Campus Party, Pentaho Day, TDC.
- Organizador Geral do Pentaho Day 2017,2015 e apoio nas edições 2013 e 2014.
- CEO da Ambiente Livre.
- Data Scientist, Instrutor e Consultor de Big Data com tecnologias abertas.

Ambiente Livre
Tecnologia

Nosso Ecossistema

Big Data e Data Science

Análise de Dados da IoT
Análise Preditiva
Processamento Distribuído
Banco de Dados Colunares

Big Data & Data Lake
Big Data Analytics
Machine Learning

Consultoria | Treinamento | Projeto

CRM e CMS

Marketing e Vendas
Fidelização
SAC e Pós-vendas
Portais de Conteúdo

Customer Relationship Management
Content Management System
Pesquisa de Mercado & SLA

Consultoria | Treinamento | Projeto

ECM e BPM

Gestão de Documentos
Gerenciamento de Mídias
Processo de Negócio
BPMN e BPMS

Enterprise Content Management
Records Management
Business Process Management

Consultoria | Treinamento | Projeto

Business Intelligence

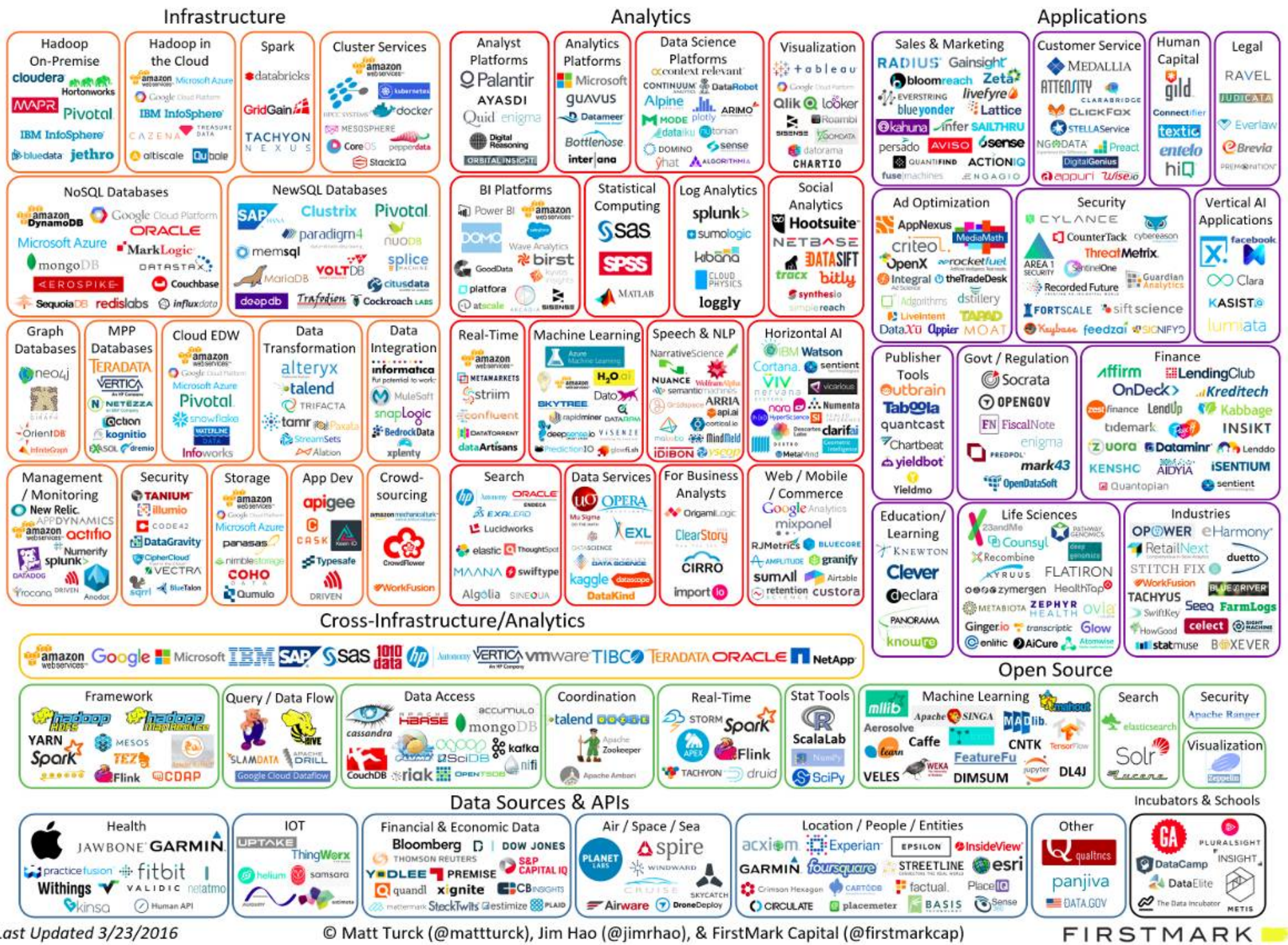
Painéis de Indicadores
Cubos de Análise
Relatórios Gerênciais
Tomada de Decisão

Business Intelligence & Analitycs
Dashboards e OLAP
Data Integration & Data Mining

Consultoria | Treinamento | Projetos

pentaho · Alfresco · hadoop · mongoDB · Cassandra · Joomla! ...because open source matters · LPI · Spark · Drupal · moodle · LIFERAY Enterprise. Open Source. For Life. · SUGARCRM · LimeSurvey · suiteCRM · MySQL · MariaDB · php · PostgreSQL the world's most advanced open source database · LibreOffice The Document Foundation · PentahoDay 2017 Curitiba - Brasil

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

Quarto paradigma da ciência

- **Empírica,** É uma maneira de adquirir conhecimento por meio de observação ou experiência direta e indireta.
- **Investigação**, Melhorar as teorias científicas para uma melhor compreensão ou previsão de fenômenos naturais. Muitas vezes impulsionado pela curiosidade.
- **Computação:** Estuda as técnicas, metodologias e instrumentos computacionais, que automatiza processos e desenvolve soluções baseadas no uso do processamento digital.
- **Baseada em dados** (**data-driven**)
  Ciência Sobre os Dados ou Ciência dos Dados

Data Science

- Campo interdisciplinar de pesquisa sobre métodos científicos, processos e sistemas para **extrair conhecimentos** ou **insights** a partir de dados em várias formas, estruturadas ou não estruturadas, semelhantes ao KDD.
- **Unificar estatísticas, análise de dados e seus métodos relacionados**, a fim de compreender e analisar fenômenos reais com dados.
- Emprega técnicas e teorias extraídas das áreas amplas de **matemática, estatística, ciência da informação e ciência da computação**, aprendizagem de máquinas, classificação, análise de cluster, mineração de dados, bancos de dados e visualização.


Software Livre
Open Source

Ambiente Livre
Tecnologia

Software Livre

- **"Software Livre"** se refere à liberdade dos usuários executarem, copiarem, distribuírem, estudarem, modificarem e aperfeiçoarem o software. São **4 tipos de liberdade**, para os usuários do software:
- **1**. A liberdade de executar o programa, para qualquer propósito.
- **2**. A liberdade de estudar como o programa funciona, e adaptá-lo para as suas necessidades. Acesso ao código-fonte é um pré-requisito para esta liberdade.
- **3**. A liberdade de redistribuir cópias de modo que você possa ajudar ao seu próximo.
- **4**. A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.

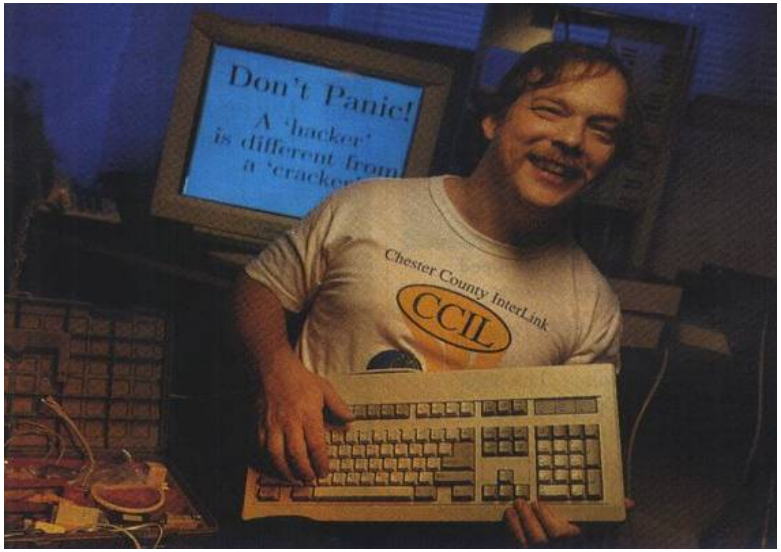
CuritibaLivre

Em Curitiba 02 de Junho!
http://rms.curitibalivre.org.br/

GPL v3 Free Software
Free as in Freedom

Open Source

- Criado pela OSI (Open Source Initiative)
- Não refere-se a software também conhecido por software livre.
- **Qualquer licença de software livre é também uma licença de código aberto (Open Source)**
- **Mas o contrário nem sempre é verdade**
- Criado por Eric Raymond e outros fundadores da OSI.



Free Software X OSI

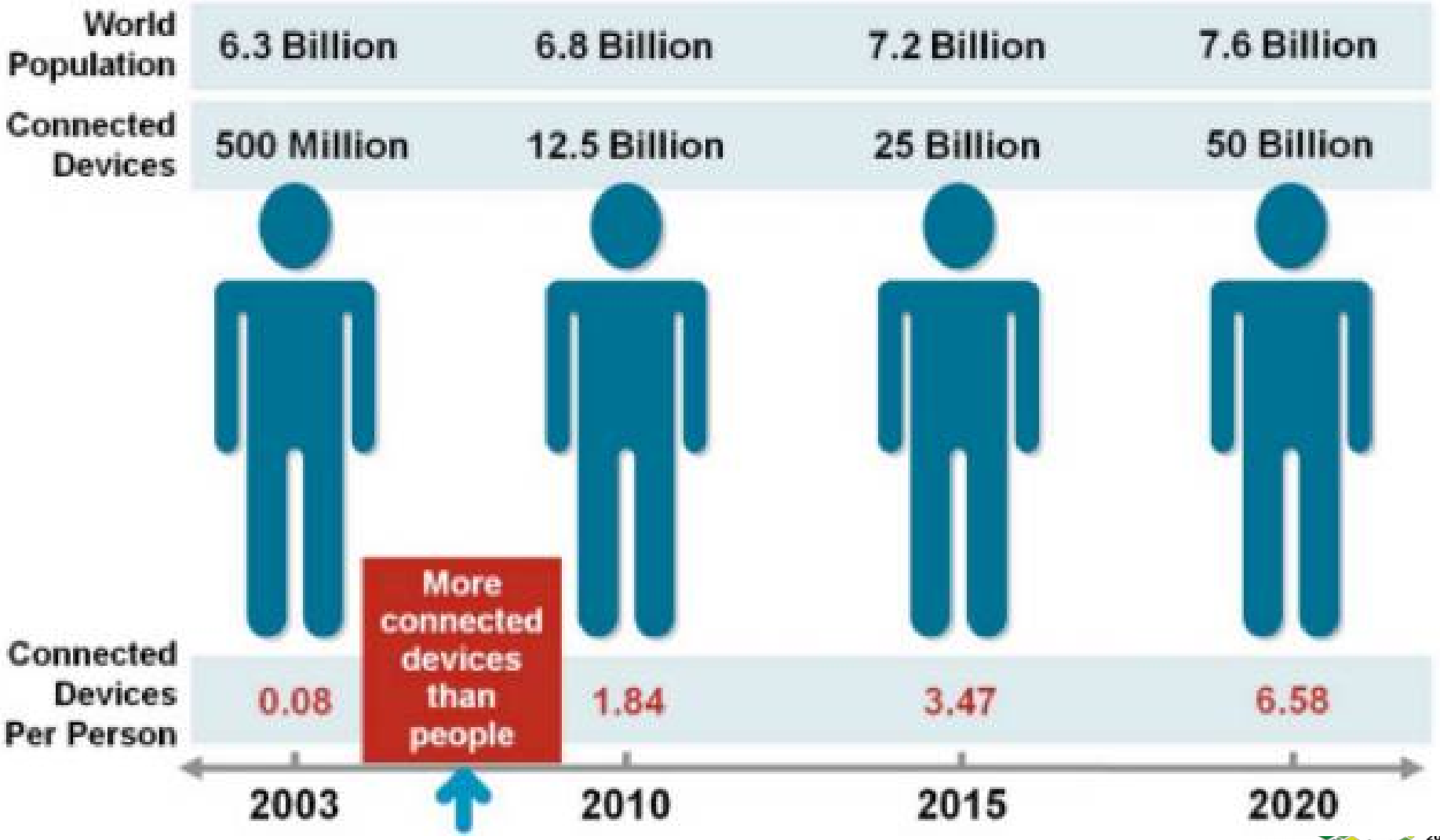
- 4 Lei da GPL
- OBRIGATORIEDADE:
  A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.

GPL Free Software
Free as in Freedom

X

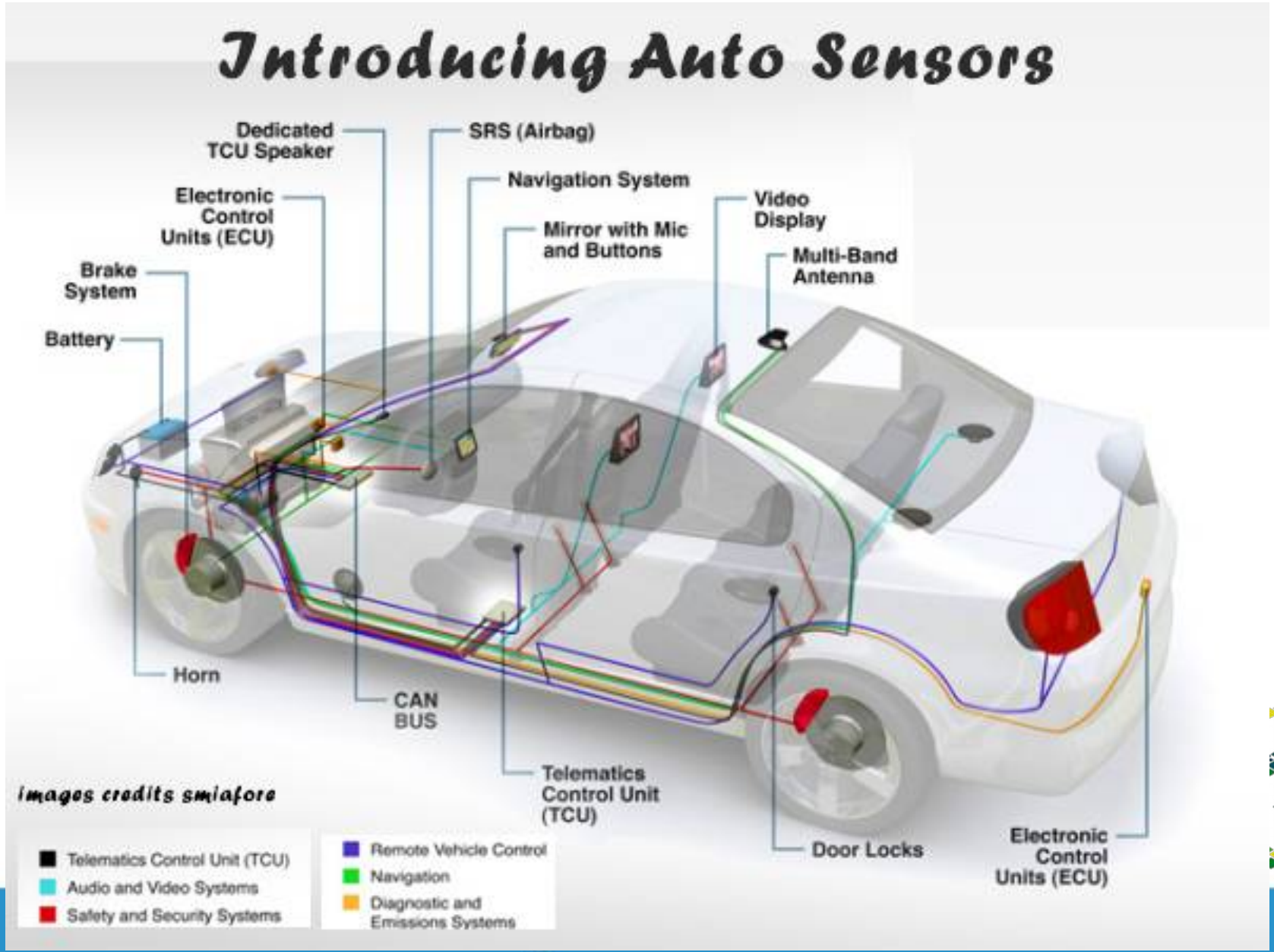
open source™

Evolução das Coisas - IOT



Ambiente Livre
Tecnologia

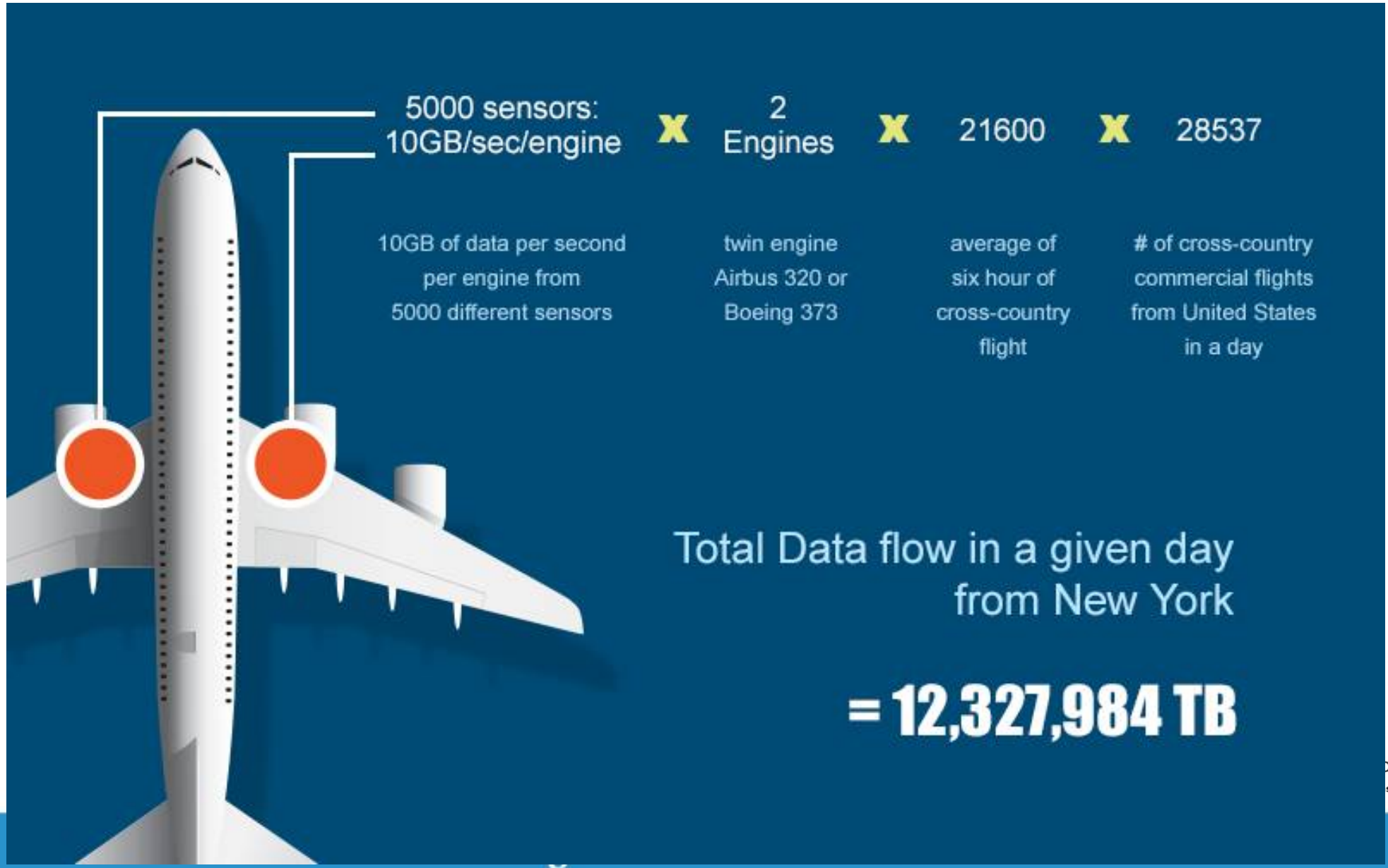
Sensores de Automóveis



PentahoDay 2017 Curitiba - Brasil

Ambiente Livre
Tecnologia

Sensores de Voo



Data Lake

- Fonte única
- Grande Volume
- Não Refinado
- Pode estar tratado.

Como era antes!

Data Mart(s)

Data Source

Ambiente Livre
Tecnologia

Arquitetura de Big Data

Data Mart(s)

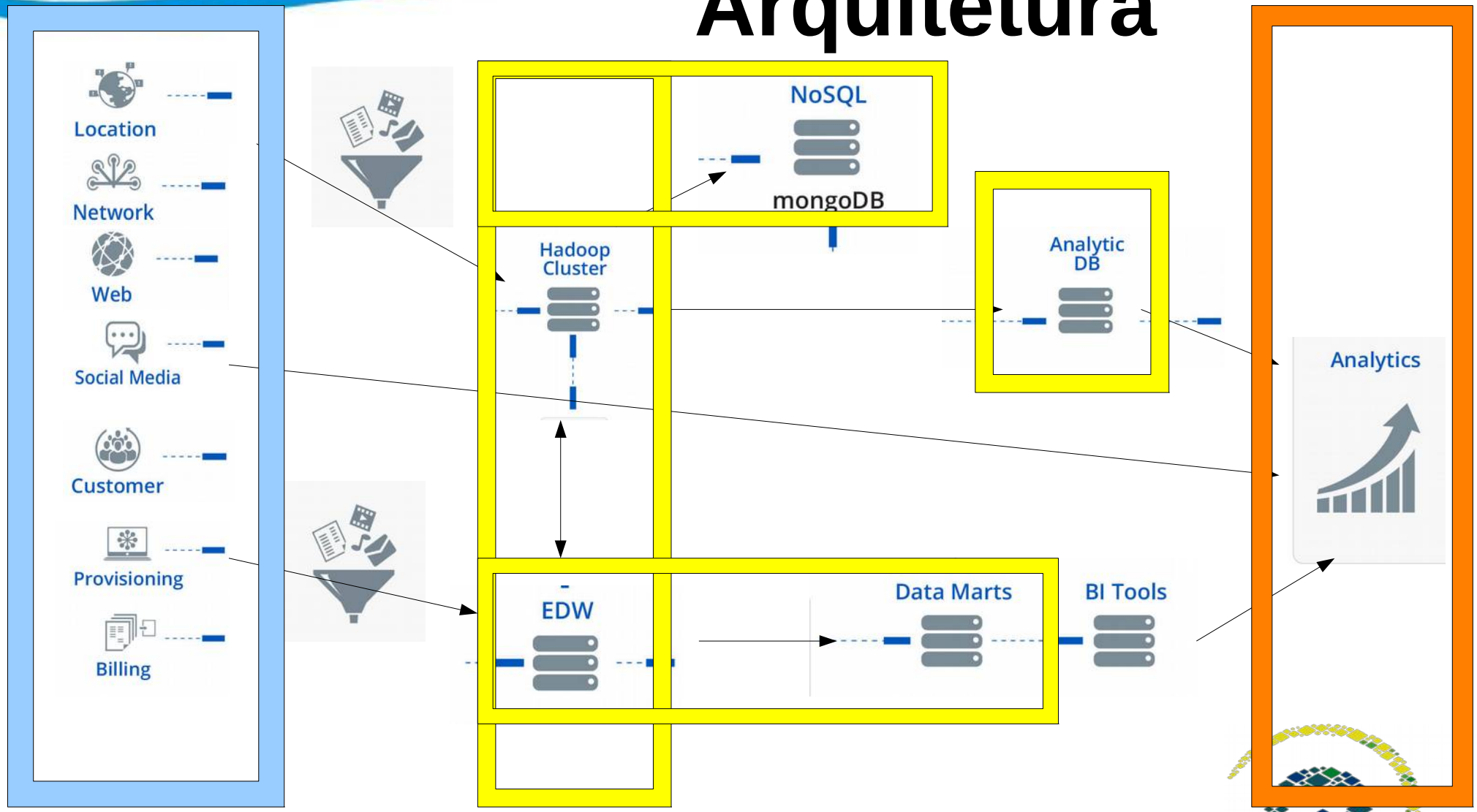
ad-hoc

Datawarehouse

Data Lake(s)

Data Source

Arquitetura



Fonte

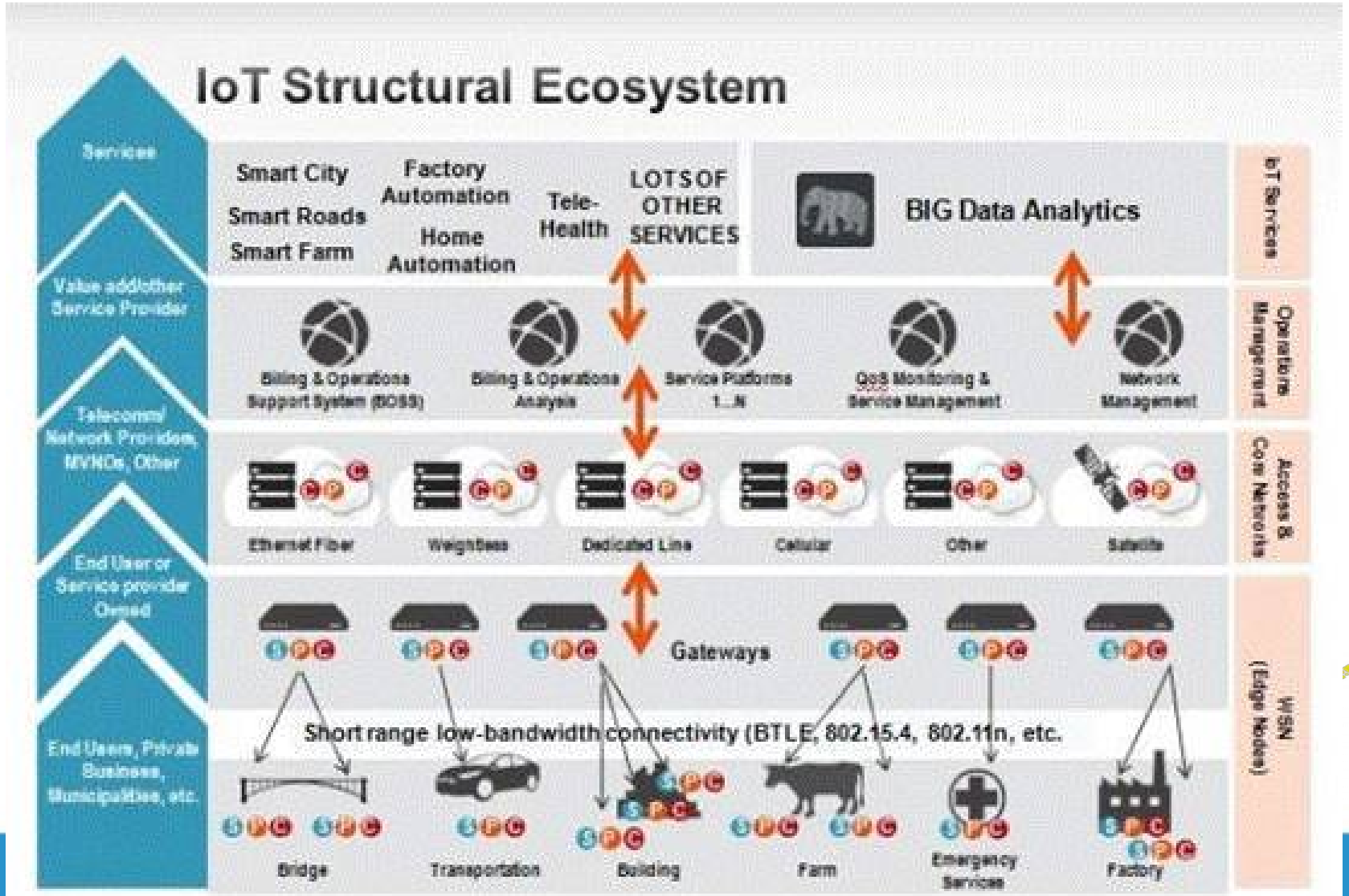
Armazenamento

Analise

Ambiente Livre
Tecnologia

Arquitetura - IoT

- U$ 4 a 11 trilhões a partir de 2025



PentahoDay 2017 Curitiba - Brasil

Captura de Dados

- Web crawler
- IoT
- Equipamentos de Redes
- Open Source (Data System) Erps, CRMs, etc
- Logs
- Etc, etc, etc

Armazenar

Armazenamento

Processar

Processamento e Integração

elasticsearch

Visualização e Analise

pentaho™
open source business intelligence™

C*TOOLS

saiku

CUTTING EDGE OPEN SOURCE ANALYTICS

D3 Data-Driven Documents

PentahoDay
2017
Curitiba - Brasil

Machine Learning

Spark MLlib

mahout

R

python™

Fundação Apache

- Data Science = Apache = Open Source
- Apache é **lider em Big Data e Data Science**!
- ~31 projetos da linha “Big Data” incluindo “Apache Hadoop” e “Spark”
-

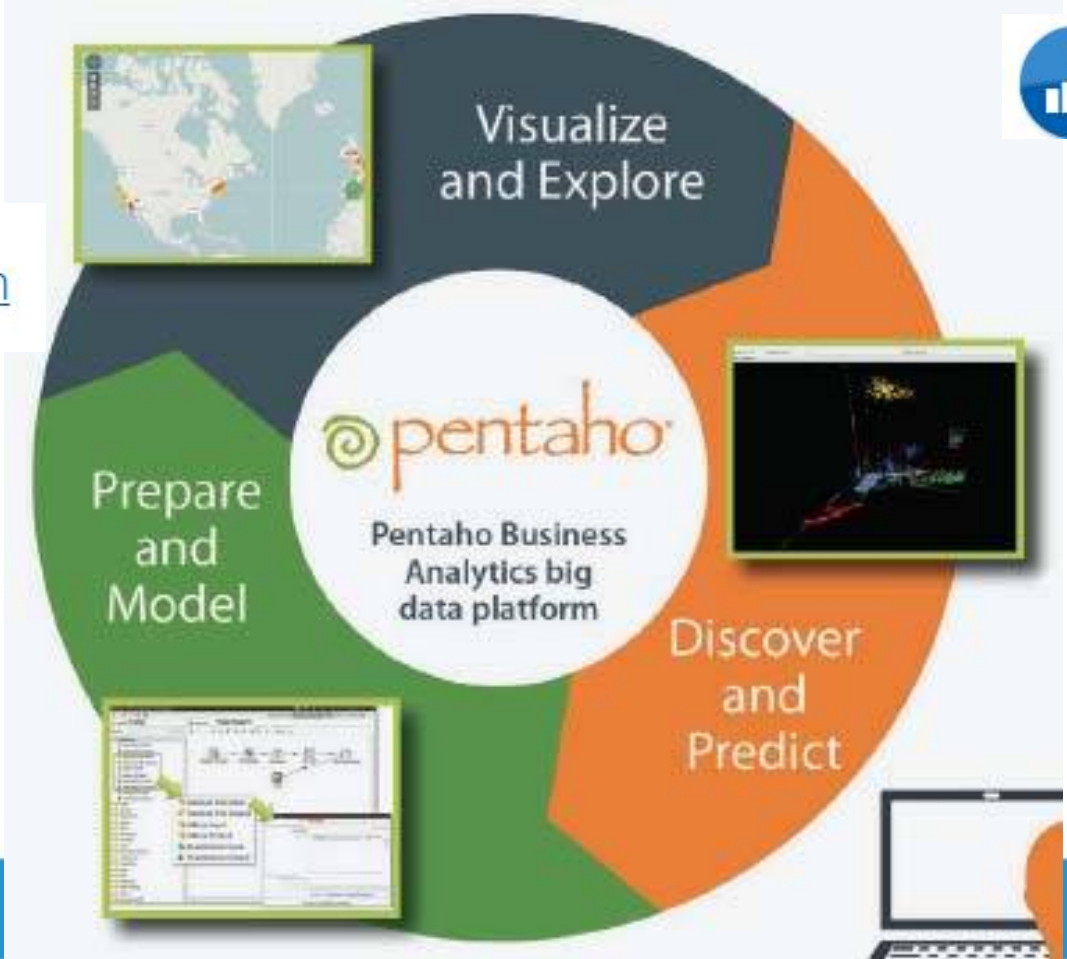
Ambiente Livre
Tecnologia

3 Pilares do Pentaho

- Plataforma abrangente para integração de dados e Business Analytics.

Business Analytics

Data Integration

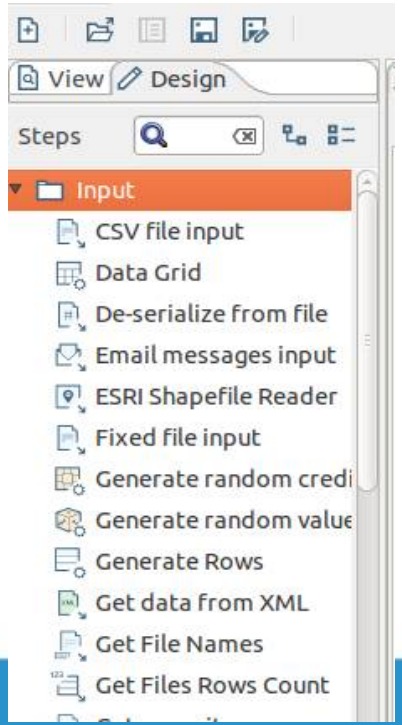


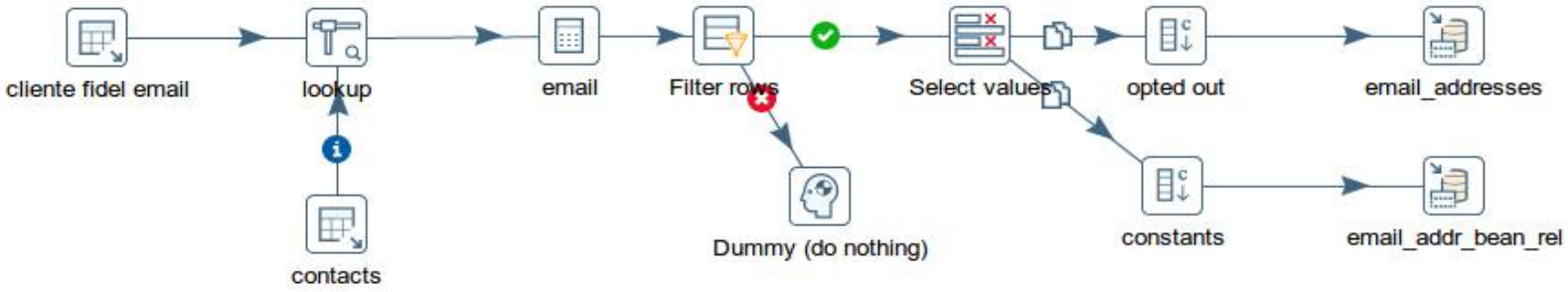
Big Data

PentahoDay 2017 Curitiba - Brasil

Pentaho Data Integration

- Processa em Paralelo (em breve em Cluster Spark)
- Acessar dados diretamente (se necessário sem DW)
- Permite publicar dados diretamente em Reports, Ad-Hoc Reports e Dasboards.
- “Programação e Fluxo Visual” com aproximadamente 350 steps diferentes





Integração ampla e adaptável de Big Data

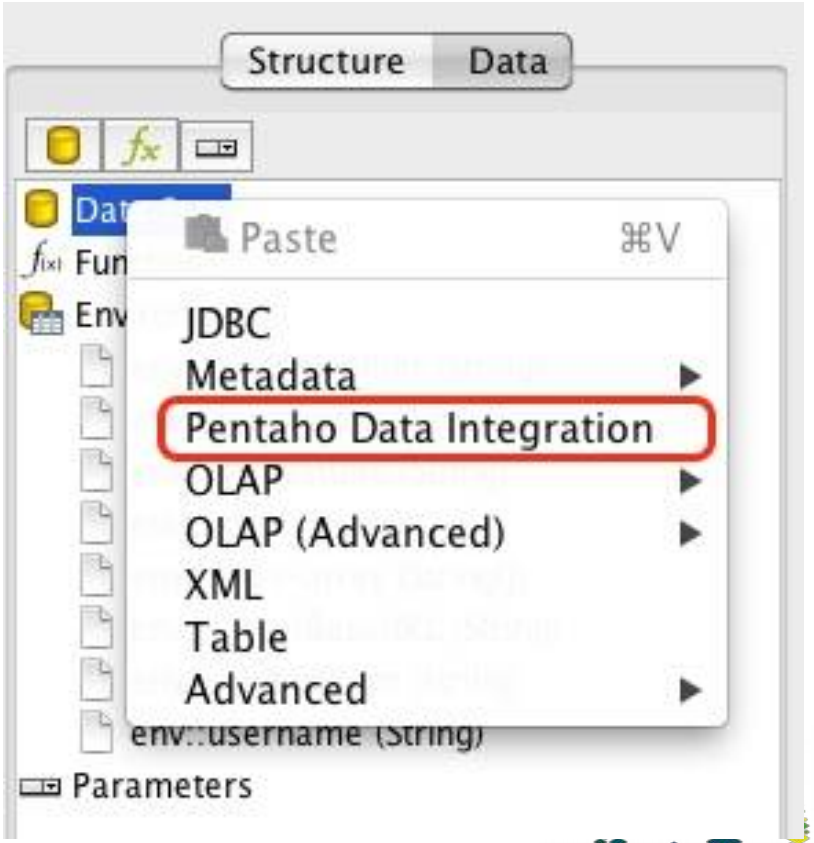
- Conexões nativas e camada adaptável de Big Data e acesso funcionalidades dos populares big data stores.
- Capacidade de acessar dados, processá-los combiná-los e consumi-los em qualquer lugar.
- Flexibilidade, isolamento das mudanças no ecossistema de dados
- Suporte a distros Hadoop
- Acessar dados para preparação via SQL no Spark e orquestrar aplicativos Spark (Scala, Java e Python)
- Integração com NoSQL stores

Pentaho Report Designer

- Visualização Web ou Embed.
- Assistente de geração de relatórios
- Amplo suporte de fonte de dados, incluindo relacionais, OLAP, XML e Pentaho Analysis, arquivos flat, objetos Java e ...
- **Big Data Reports (integra-se com PDI)**

ETL como Data Source

- O data source do report é um ETL.
- Isso muda tudo!



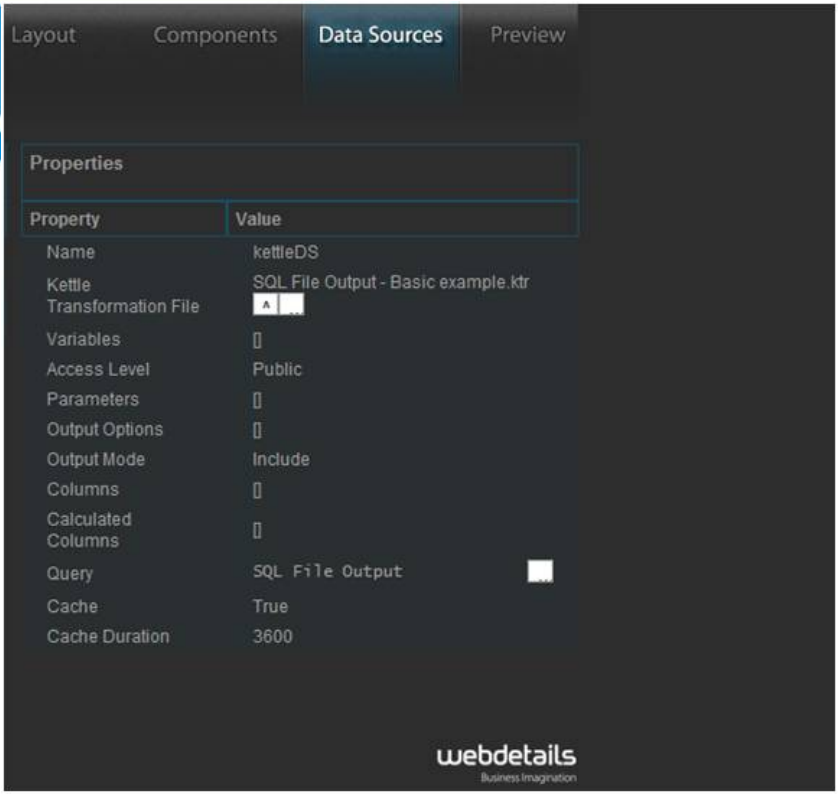
Exemplo de dados do Twitter Report

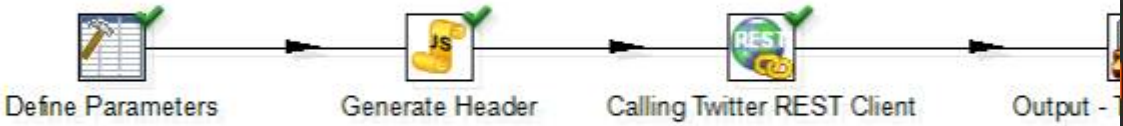
- Libere na API acesso
- Crie seu ETL no PDI (Pentaho Data Integration)
- Defina onde quer os dados (database, hadoop, Report ou dashboard)

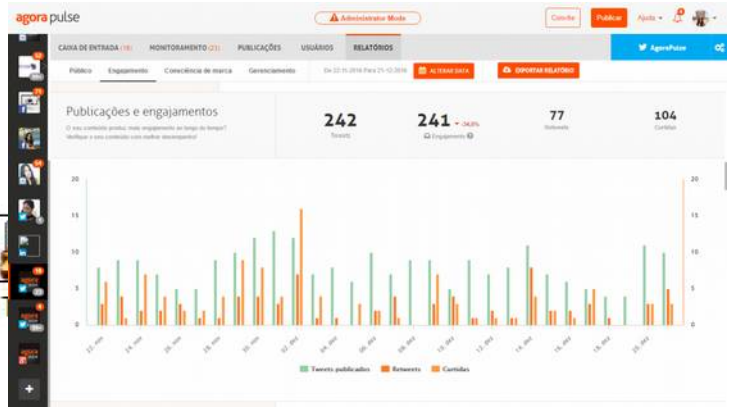
PentahoDay 2017 Curitiba - Brasil

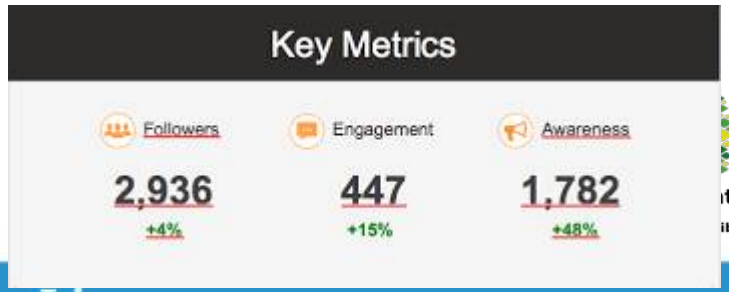
Dashboards ETL

- Dashboards permiter integração com ETL

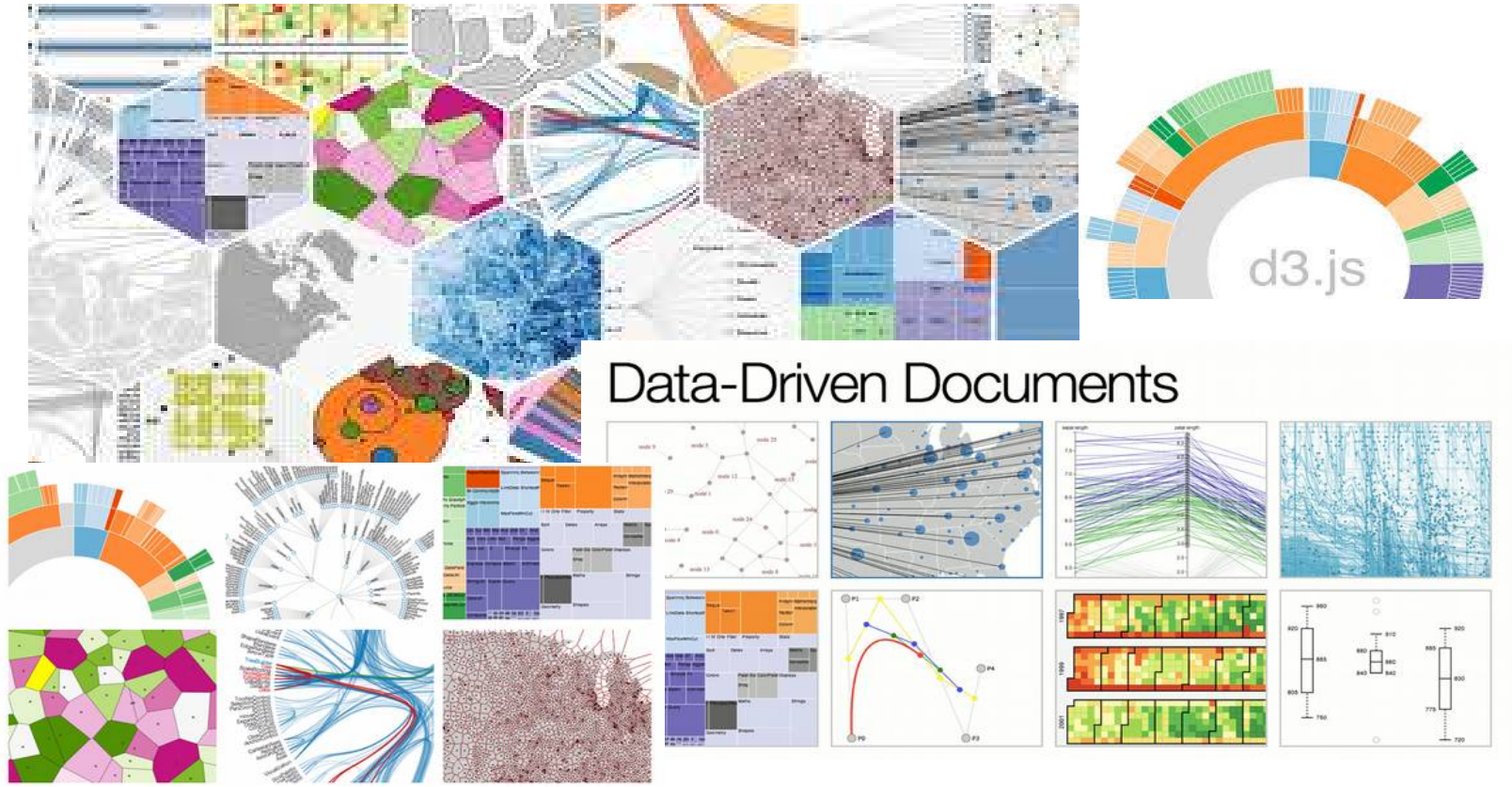








ETL para datasets D3.js



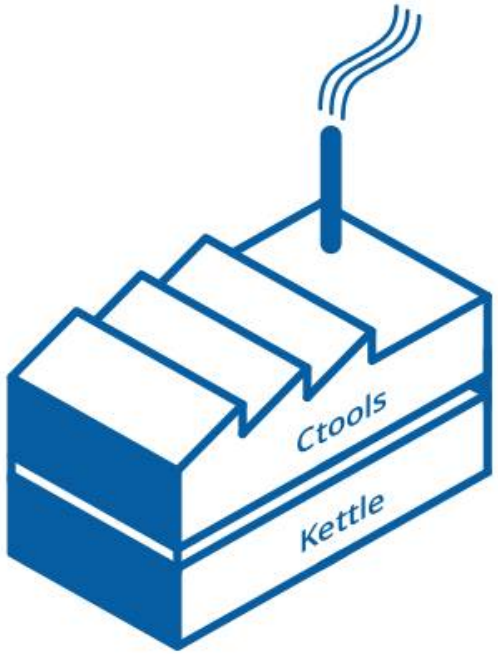
http://romsson.github.io/dragit/example/nations.html

https://bl.ocks.org/mbostock/1136236

http://bl.ocks.org/brattonc/5e5ce9beee483220e2f6

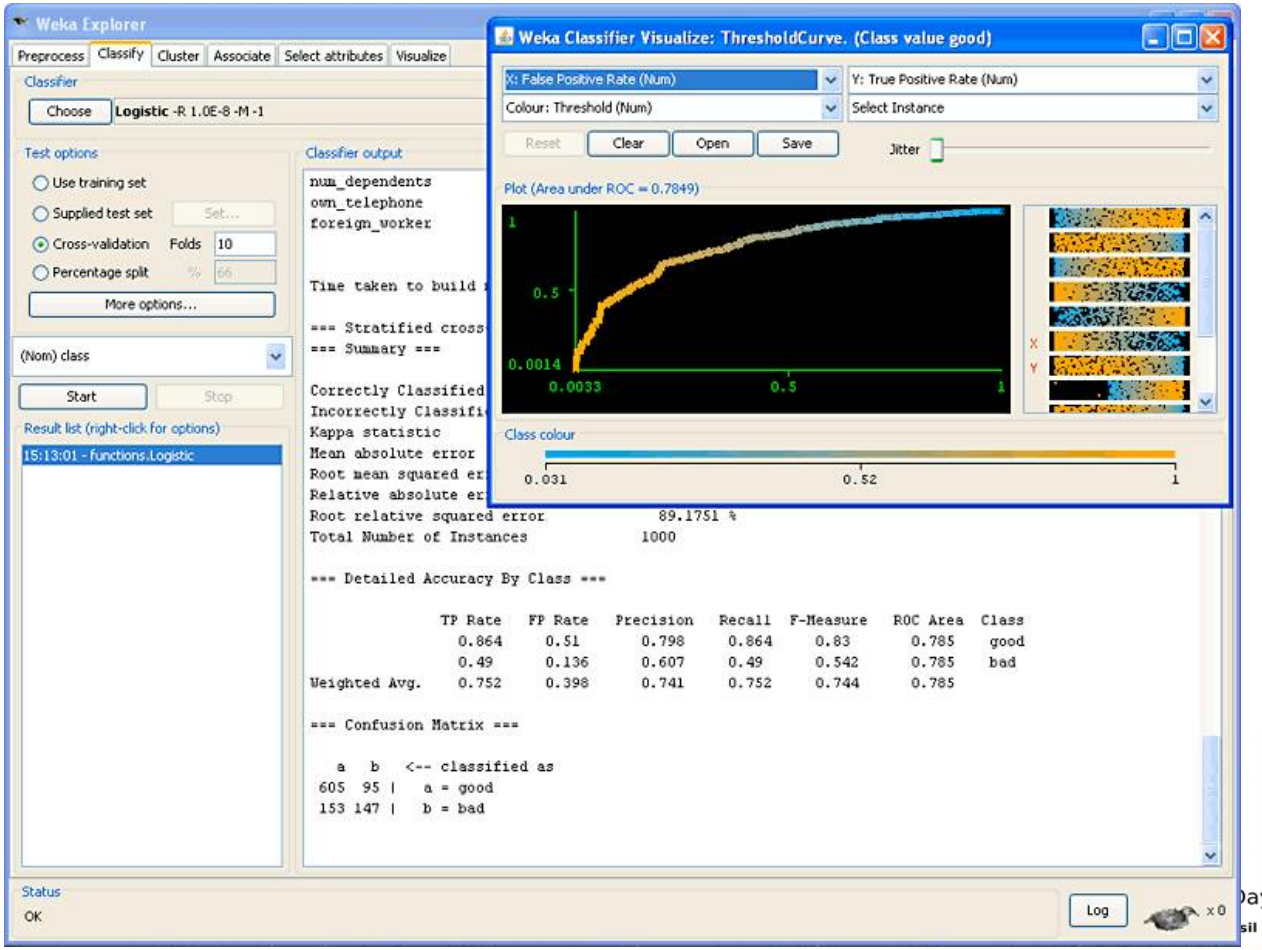
Pentaho Sparkl

- Framework que usa o PDI como “fonte”
- App Builder que permite desenvolver plugins de Big Data Analytics e outros em alguns passos.
- Menus = Dados
- Campos = meta**Dados**
- Botão = Dispara **Serviço**
- Filtros = Lista **Dados**
- **Todos mais faça JS/Jquery :)**



Pentaho Data Mining

- Solução completa para Machine Learning
- Aprox. 79 Algorítimos
  - Classificação
  - Associação
  - Cluster



Comunidade Brasileira

PentahoDay
2017
Curitiba - Brasil

Comunidade Brasileira

- Maior comunidade do Mundo!
- Lista de Discussão com + de 1900 membros
- Organiza a 7 anos o Pentaho Day Brasil
- Composta por desenvolvedores, usuários , empresas e acadêmia.
- Utilizado em mais de 185 países.
- +10.000 Produtos desenvolvidos sobre a plataforma Pentaho.
- + 4 milhões de Downloads
- Em 2015 +- 60.000 downloads dia

	Country	Android	BSD	Linux	Macintosh	Solaris	Unknown	Windows	Total
1.	United States	0%	0%	7%	16%	0%	15%	62%	50,213
2.	Brazil	0%	0%	15%	5%	0%	3%	77%	41,115
3.	China	5%	0%	3%	4%	0%	2%	86%	39,910
4.	Germany	0%	0%	7%	7%	0%	45%	41%	20,695

oDay
Brasil

Open Source gera valor

- Facebook vende software? Não mas entrega muita tecnologia open source assim como milhares de outras startup. Exemplo Hive.

Home › Intel anuncia investimento de US$ 740 milhões na Cloudera e fortalece Big Data

Intel anuncia investimento de US$ 740 milhões na Cloudera e fortalece Big Data

Por Redação em | 31.03.2014 às 17h30

Recomendar 2 Tweetar 14 +1 0 Share http://canalte.ch/SEKG

A Intel, a maior fabricante de chips do mundo, anunciou neste último fim de semana o investimento de US$ 740 milhões em 18% das ações da Cloudera, empresa especializada em

Dificuldades ou Desculpas criadas por “vendos”

- Como vai gerenciar Schedulers ?
- Como vai gerenciar Segurança ?
- Como vai gerenciar o Cluster ?

Como ? Como ? Como?

- cron
- chmod 600
- Shell script
- Open Source

Data Scientist Nutela

Data Scientist Raiz

Diferenciais Reais mas não impeditivos

- Interface
- Aceleração do Trabalho
- BI Self Service – Será mesmo ?
- Suporte do Desenvolvedor

Dificuldades Reais

- Alto investimento em capital intelectual das pessoas
- Encontrar pessoas com perfil “hacker e pesquisador”
- Tempo
- **Persistência**

Acontecendo no mercado

- Compram Player de Mercado...
- Montamos Cluster na Amazon, Azure, Azure
- Uso o Framework da Nuvem
- O custo sobe.. a empresa cresce.. e crise vem... o dólar sobe...!
- Começo a mesclar usando Open Source
- Startups! Começam ao Contrário! Open Sorce sempre primeiro.

Minhas Perguntas aos Grandes

- Sei que você usa arquitetura “mesclada”, mas é possível fazer 100% Open Source?
- Sim recebidos!

Ambiente Livre
Tecnologia

Data Science 100% Open Source

SIM by

The Apache Software Foundation

Contatos

- marcio @ ambientelivre.com.br
- http://twitter.com/ambientelivre
- @ambientelivre
- @marciojvieira
- Blog: blogs.ambientelivre.com.br/marcio
- Facebook/ambientelivre