

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.



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 Gerenciais
Tomada de Decisão

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

Consultoria | Treinamento | Projetos



Big Data Landscape 2016 (Version 3.0)

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Duoble

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Duoble

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITALINSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, Bottlenose, interana

Data Science Platforms
context relevant, CONTINUUM, DataRobot, Alpine, MODE, plotly, dataiku, ARIMO, dataiku, ARIMO, dataiku, ARIMO

Visualization
tableau, Google Cloud Platform, Qlik, looker, Roambi, Sisense, Datorama, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fusemachines, ENGAGIO

Customer Service
MEDALLIA, ATTENSI, CLARABridge, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appurri, wiseio

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMATION

NoSQL Databases
amazon, Google Cloud Platform, Microsoft Azure, mongoDB, <EROSPIKE>, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP, Clustrix, Pivotal, paradigm4, memsql, nuodb, splice, VOLTDB, citusdata, MariaDB, Traefik, Cockroach LABS

BI Platforms
Power BI, amazon, Wave Analytics, GoodData, birist, platform, atscale, Qlik, Tableau

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, cloud physics, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthetio, simple reach

Ad Optimization
AppNexus, MediaMath, critico, OpenX, rocketHub, Integral, theTradeDesk, Adgithers, dsillery, DataXu, Appier, MOAT

Security
CYLANCE, CounterTack, cybereason, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, signifyd

Vertical AI Applications
facebook, Clara, KASIST, lumina

Graph Databases
neo4j, SHAF4, OrientDB, InfoGraph

MPP Databases
TERADATA, VERTICA, Netezza, action, cognito, SASOL, dremio

Cloud EDW
amazon, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, MATRIUM, InfoWorks

Data Transformation
alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

Real-Time
amazon, METAMARKETS, stream, confluent, DATATORRENT, dataArtisans

Machine Learning
Amazon Machine Learning, H2O.ai, SKYTREE, rapidminer, DATATORRENT, deepcentro, VISEER, predictionIO, gliflash

Speech & NLP
NarrativeScience, NUANCE, semantic, ARRIA, nora, HyperSense, contextual io, mind.io, IDIBON, yscop

Horizontal AI
IBM Watson, Cortana, sentiment, vicarious, numenta, darifai, DETER, MetaMind

Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, EN, FiscalNote, enigma, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, Kabbage, tidemark, INSIGHT, ZUORA, Dataminr, Lenddo, KENSHO, AIDYA, ISENTIUM, Quantopian

Management / Monitoring
New Relic, APPDYNAMICS, amazon, actifio, Numerify, splunk, BRADDOG, DRIVEN, Anodot

Security
TANUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO, Qumulo

App Dev
apigee, CASK, Typesafe, DRIVEN

Crowd-sourcing
amazon, mechanical turk, CrowdFlower, WorkFusion

Search
hp, ORACLE, EXALGO, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

Data Services
UO OPERA, Mo Signa, EXL, DATA SCIENCE, kaggle, dataScope, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPITUDE, granify, sumall, Airtable, retention, custora

Education / Learning
KNEWTON, Clever, Declara, PANORAMA, knowre

Life Sciences
23andMe, Counsyl, Recombin, KYRUS, FLATIRON, oozymergen, HealthTop, METABIOTA, ZEPHYR, HEALTH, OVI, Gingerio, transcriptic, Glow, Centic, AiCure, Atomwise

Industries
OPPOWER, eHarmony, RetailNext, STITCH FIX, WorkFusion, TACHYUS, Swiftkey, Seeq, FarmLogs, collect, BBOXER

Cross-Infrastructure/Analytics

amazon, Google, Microsoft, IBM, SAP, sas, data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop, YARN, Spark, MESOS, TEZ, Flink, CDAP

Query / Data Flow
SLAMDATA, DRILL, Google Cloud Dataflow

Data Access
HBASE, accumulo, mongoDB, cassandra, kafka, CouchDB, riak, OPENFDB, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
ScalaLab, SciPy

Machine Learning
mlilb, Aerosolve, Caffe, CNTK, TensorFlow, WEKA, FeatureFu, DIMSUM, suppler, DL4J

Search
elasticsearch, Solr

Security
Apache Ranger, Visualization, Espell

Data Sources & APIs

Health
Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YDLEE, PREMISE, quandl, xignite, CBINIGHTS, metamark, StockTwits, Gestimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, Airware, DroneDeploy, SKYCATCH

Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTO, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

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.



An Emperor penguin stands on a vast, flat expanse of snow and ice. The penguin is white with a black head and back, and a distinctive yellow patch on its neck. It is facing right. In the background, there are low, snow-covered mountains under a pale, overcast sky.

Software Libre

Open Source

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.



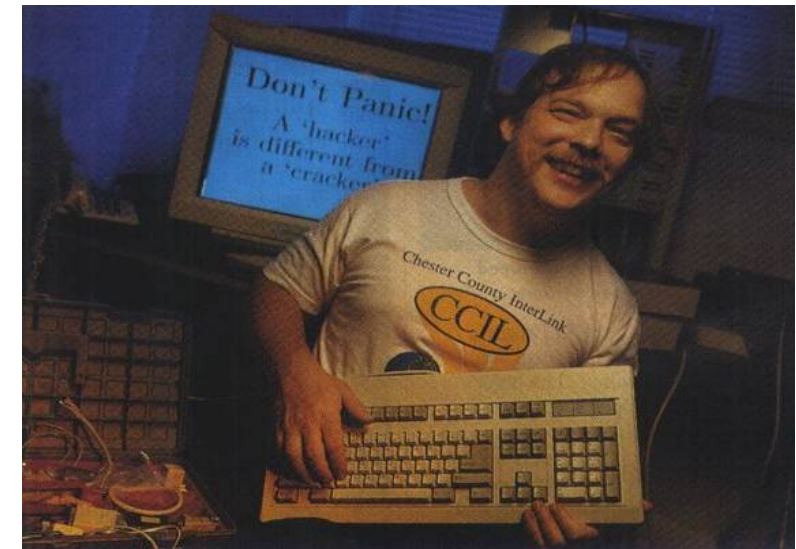
Em Curitiba 02 de Junho!
<http://rms.curitibalivre.org.br/>



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.

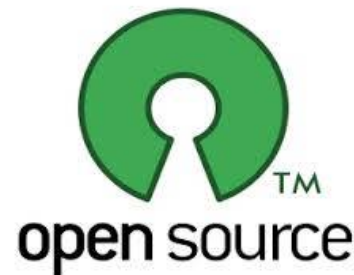


Free Software X OSI

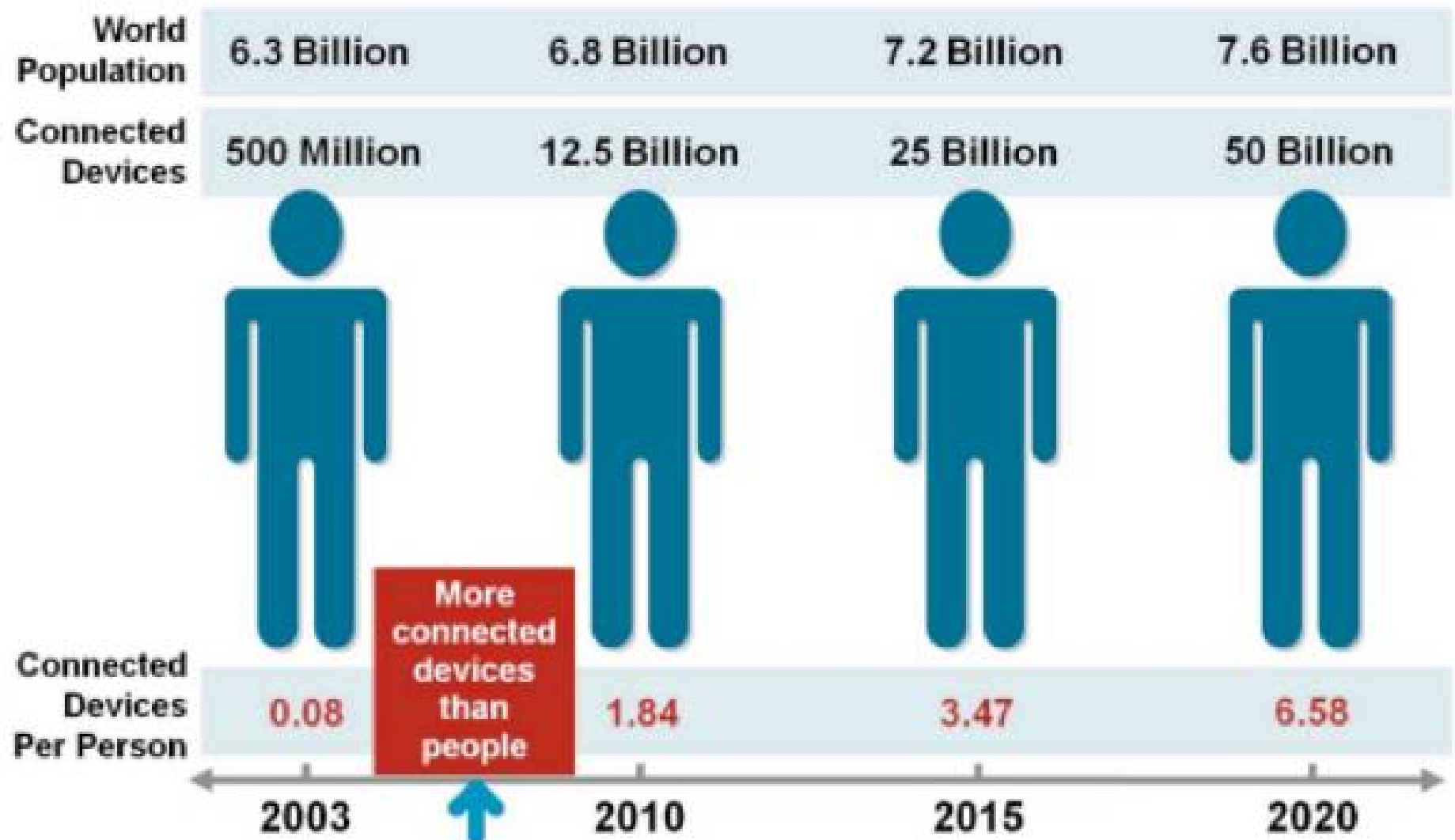
- 4 Lei da GPL
- OBRIGATORIEDADE:
A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie.



X

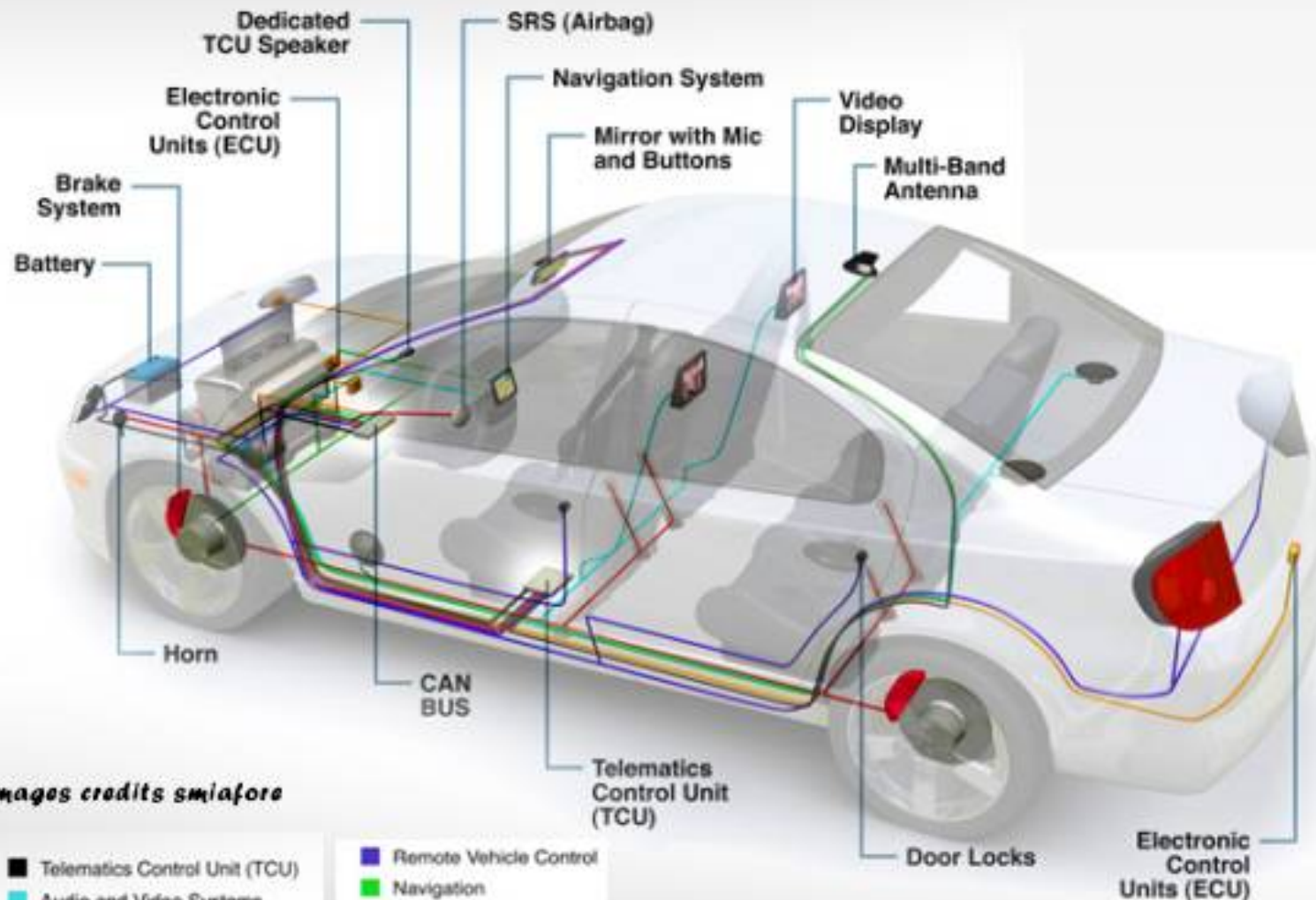


Evolução das Coisas - IOT



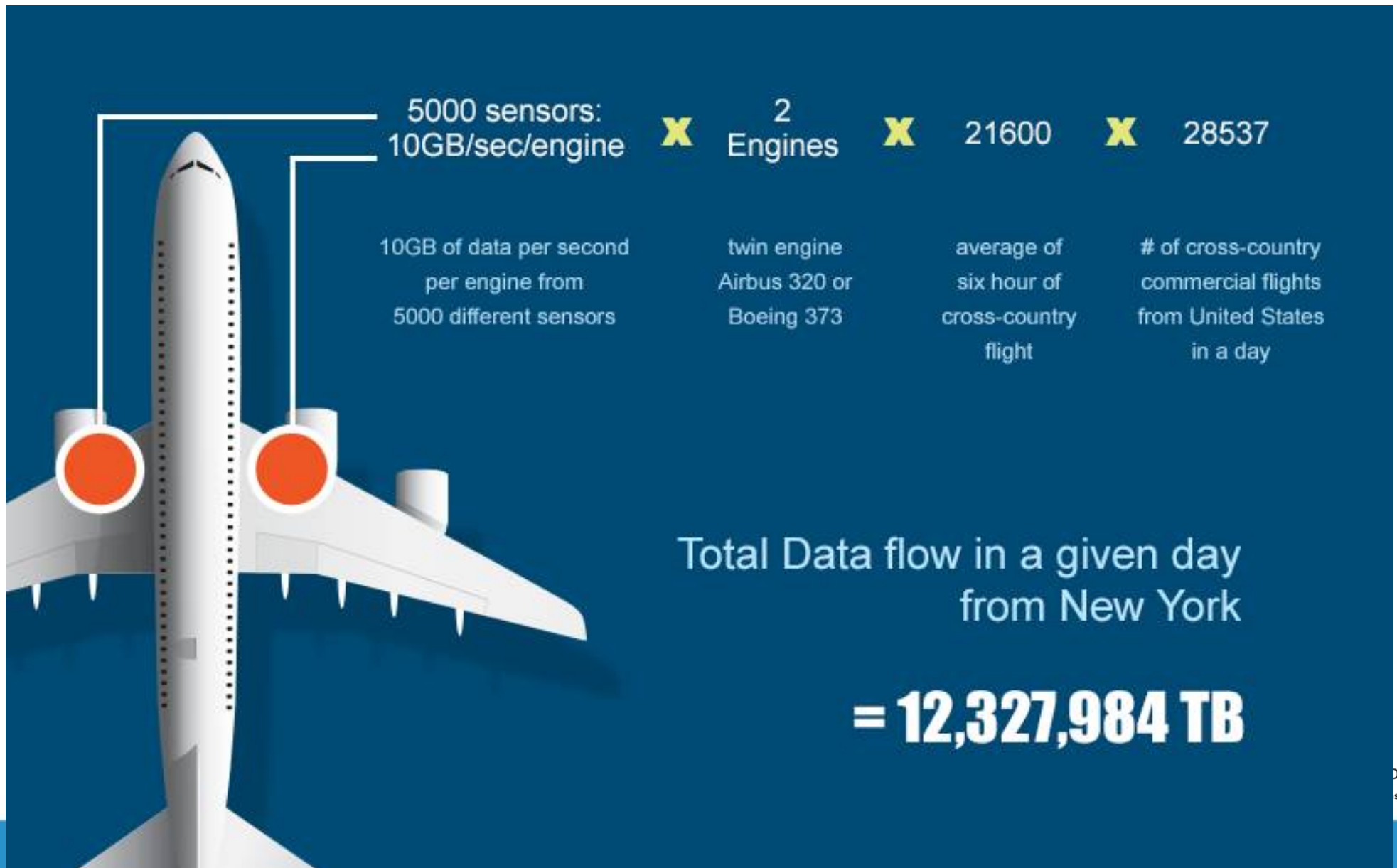
Sensores de Automóveis

Introducing Auto Sensors



Images credits smiafore

Sensores de Voo



Data Lake

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



Como era antes!

Data Mart(s)



Data Source



Arquitetura de Big Data

Data Mart(s)



ad-hoc



Datawarehouse



Data Lake(s)



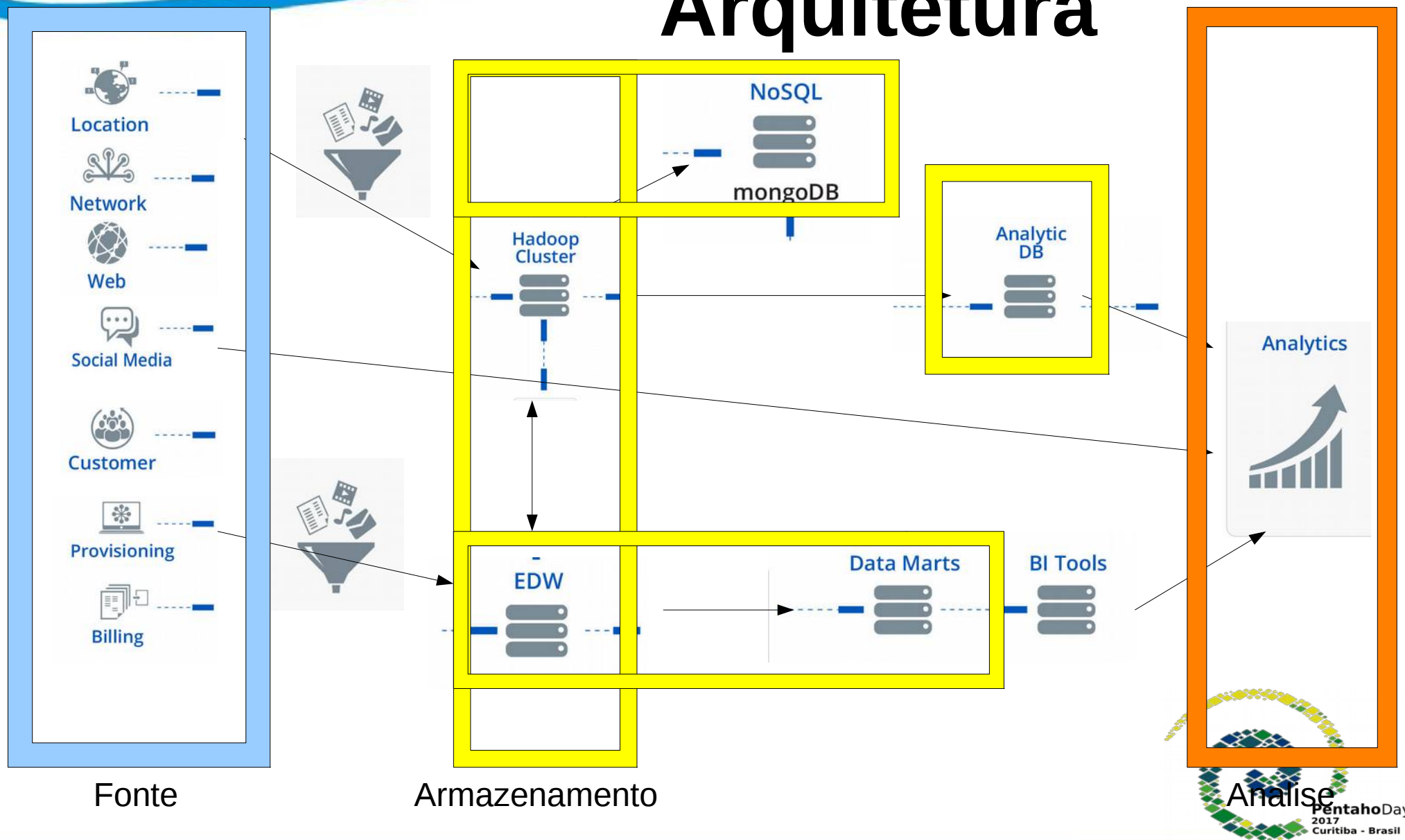
Data Source

Tecnologia



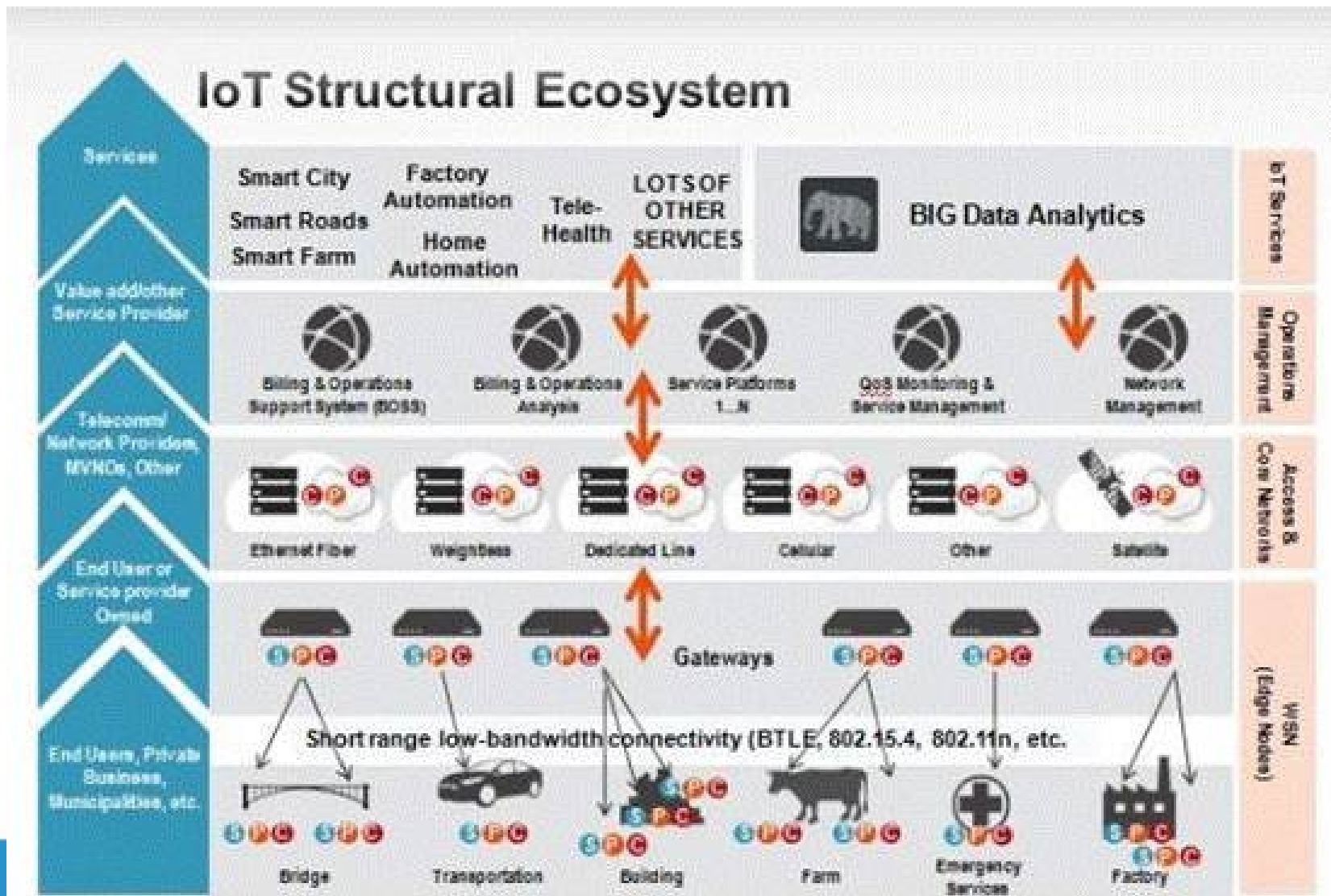
PentahoDay
2017
Curitiba - Brasil

Arquitetura



Arquitetura - IoT

- U\$ 4 a 11 trilhões a partir de 2025



Captura de Dados

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



AUTOMOTIVE
GRADE LINUX



ZABBIX



Armazenar



Armazenamento

Infinispan



Processar



Processamento e Integração



elasticsearch



Visualização e Analise



CUTTING EDGE OPEN SOURCE ANALYTICS



Machine Learning



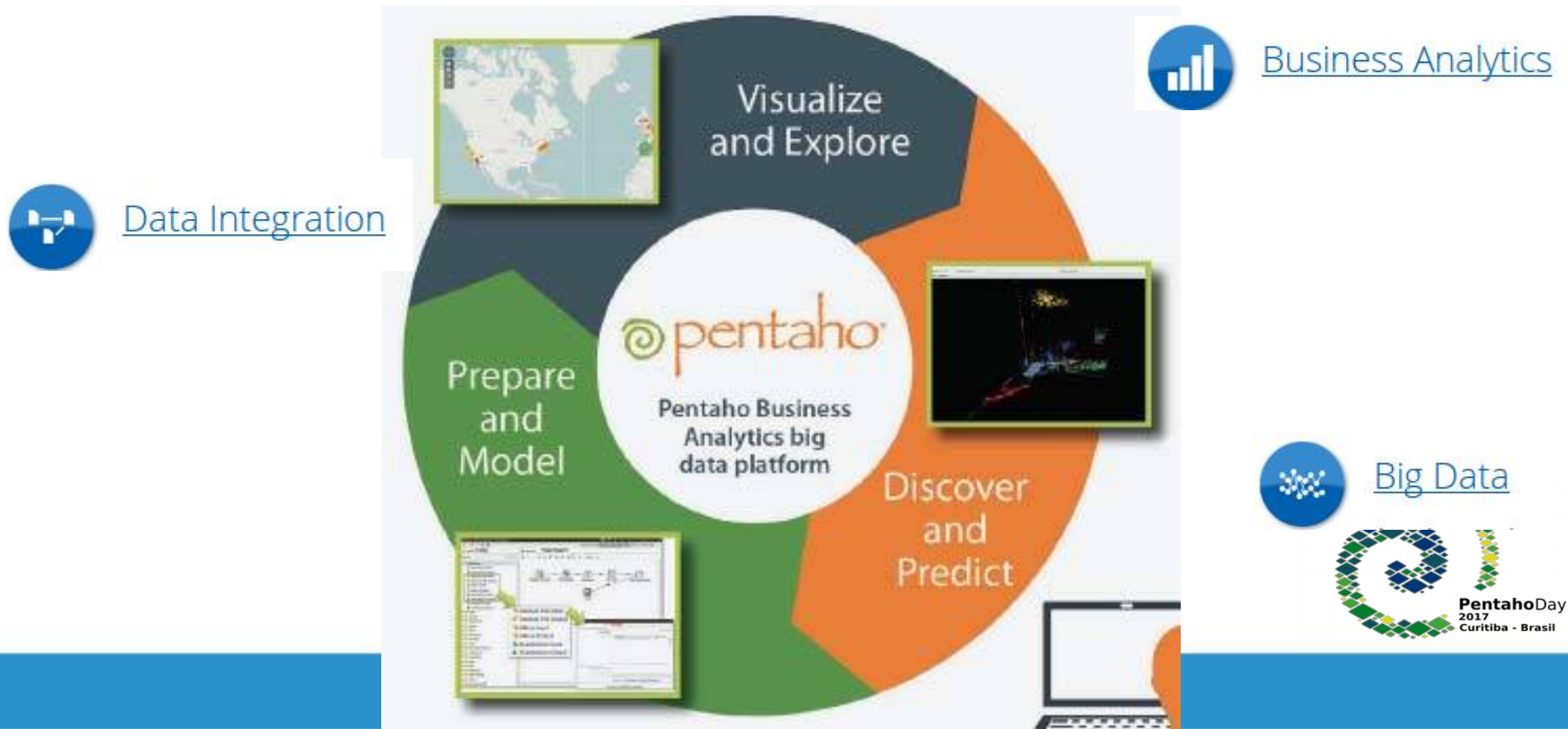
Fundação Apache

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



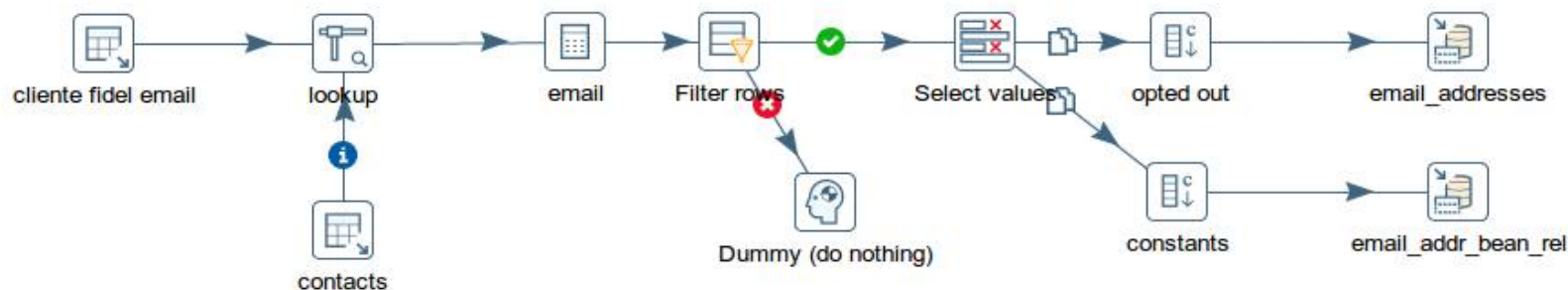
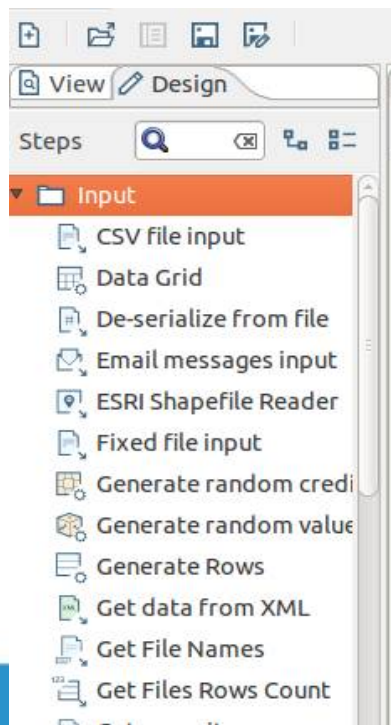
3 Pilares do Pentaho

- Plataforma abrangente para integração de dados e Business Analytics.



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 Dashboards.
- “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



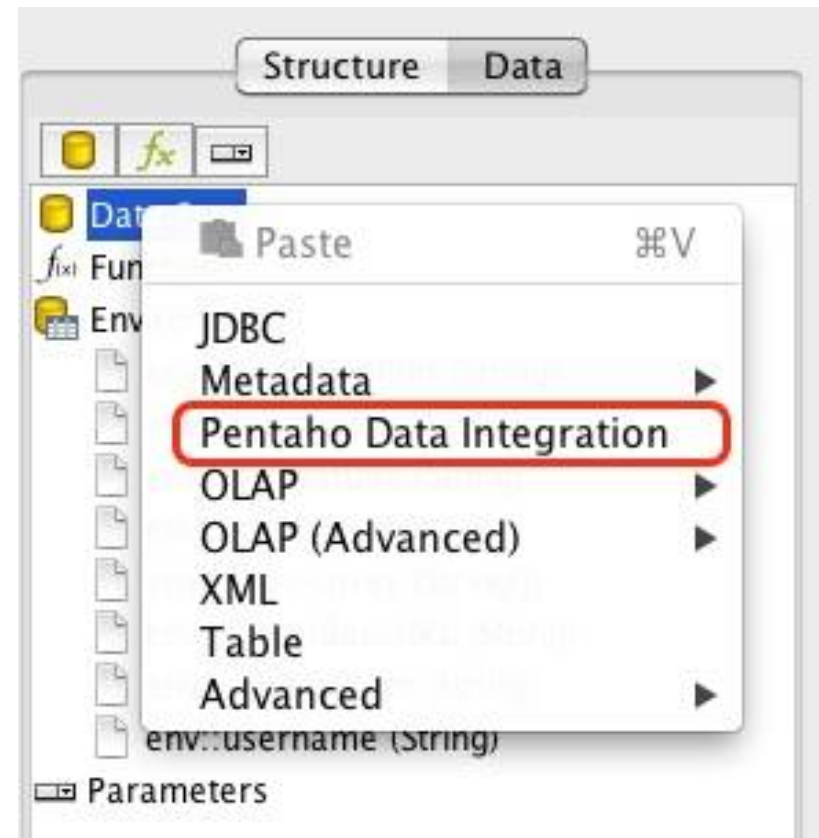
Pentaho Report Designer

- 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!



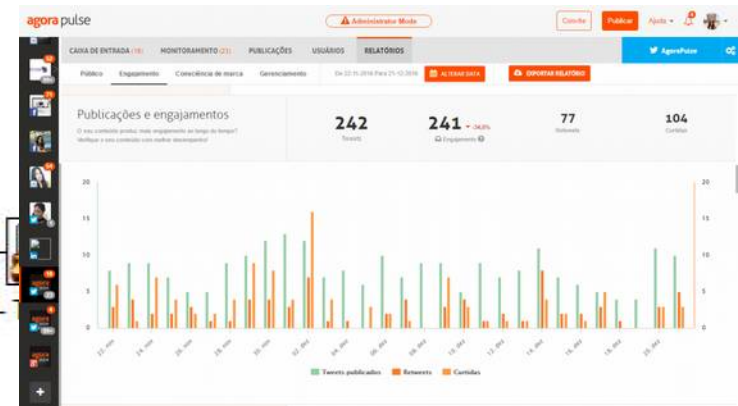
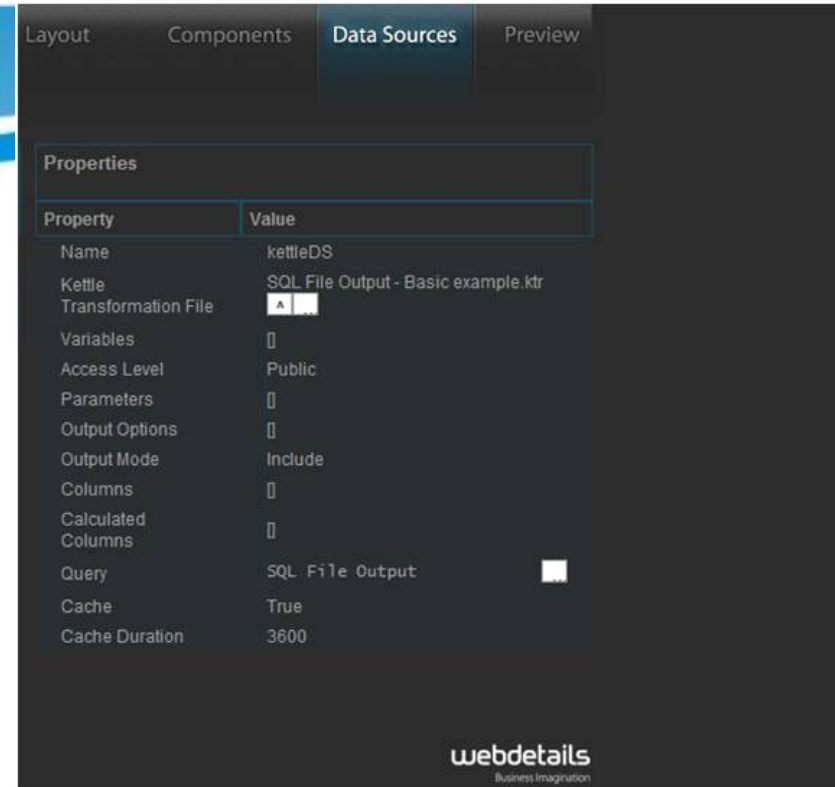
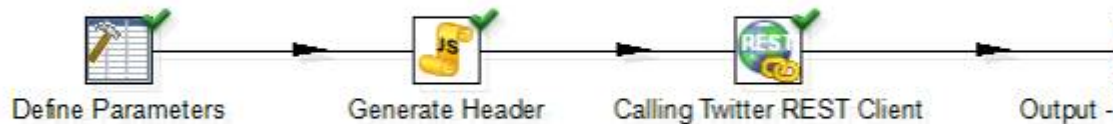
Exemplo de dados do Twitter Report

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

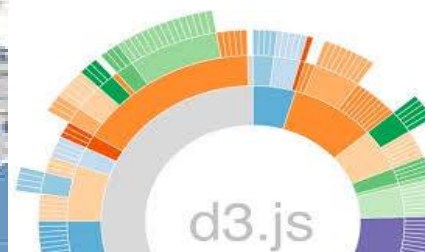
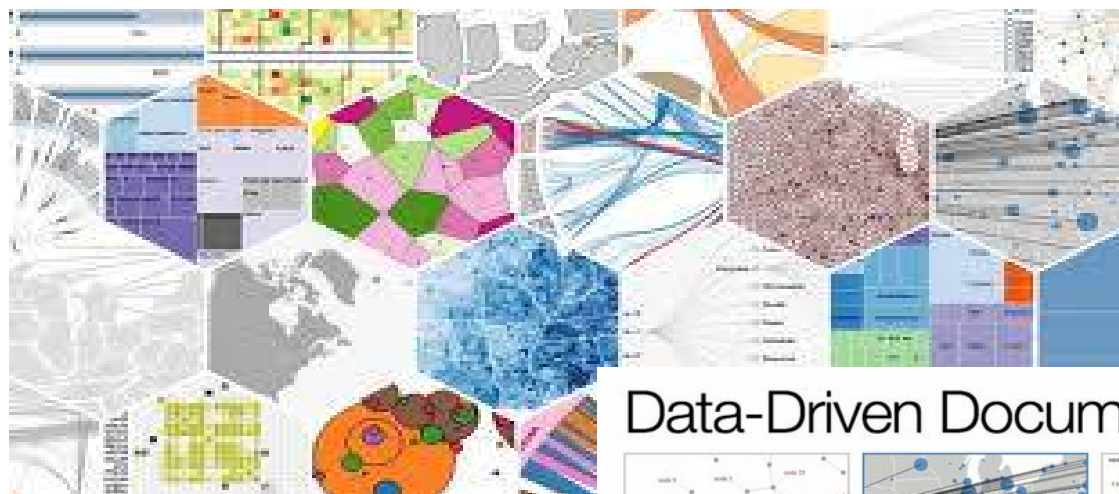


Dashboards ETL

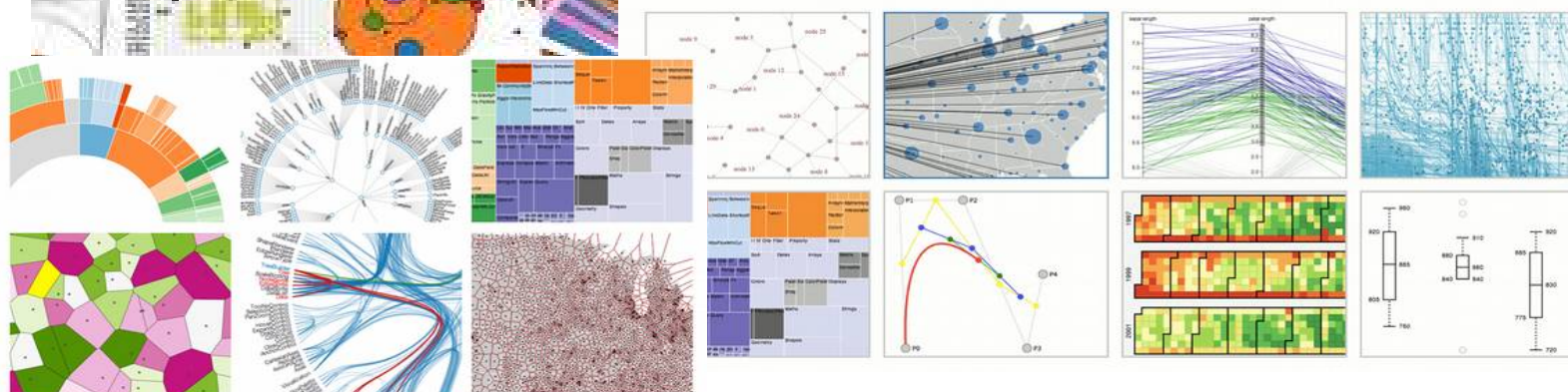
- Dashboards permitir integração com ETL



ETL para datasets D3.js



Data-Driven Documents



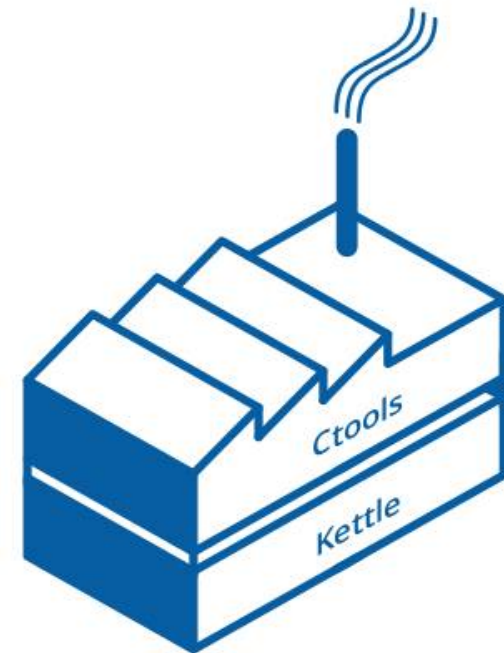
<http://romsson.github.io/dragit/example/nations.html>

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

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

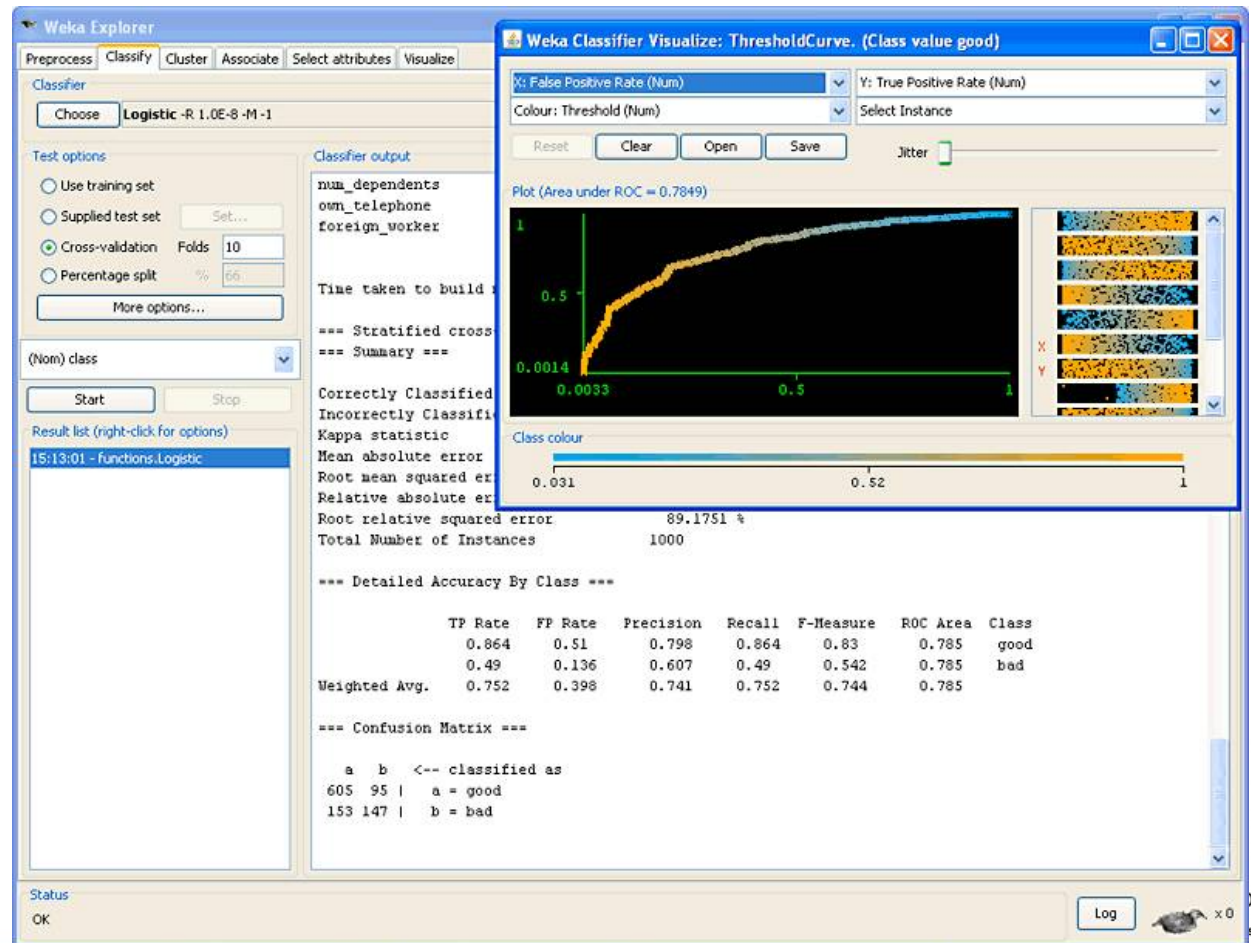
Pentaho Sparkl

- 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 Algoritmos
 - Classificação
 - Associação
 - Cluster



Comunidade Brasileira



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 academia.
- 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

Open Source gera valor

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



Dificuldades ou Desculpas criadas por “vendos”

- Como vai gerenciar Schedulers ? • cron
- Como vai gerenciar Segurança ? • chmod 600
- Como vai gerenciar o Cluster ? • Shell script
- Como ? Como ? Como? • 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 Source sempre primeiro.



Minhas Perguntas aos Grandes

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

NETSHOES

luizalabs
magazineluiza
vem ser feliz



vivo
Telefonica



PETROBRAS



PentahoDay
2017
Curitiba - Brasil

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