

Big Data

Programa

- 18/11/2020 – Conceitos em Big Data, Data Science em IoT
- 25/11/2020 – Introdução ao Spark e PySpark
- 02/12/2020 - Aprendendo Apache Spark com PySpark e Databricks
- 09/12/2020 - Structured Streaming no PySpark
- 16/12/2020 – Modelagem Preditiva no PySpark

Parte I – Definindo Big Data

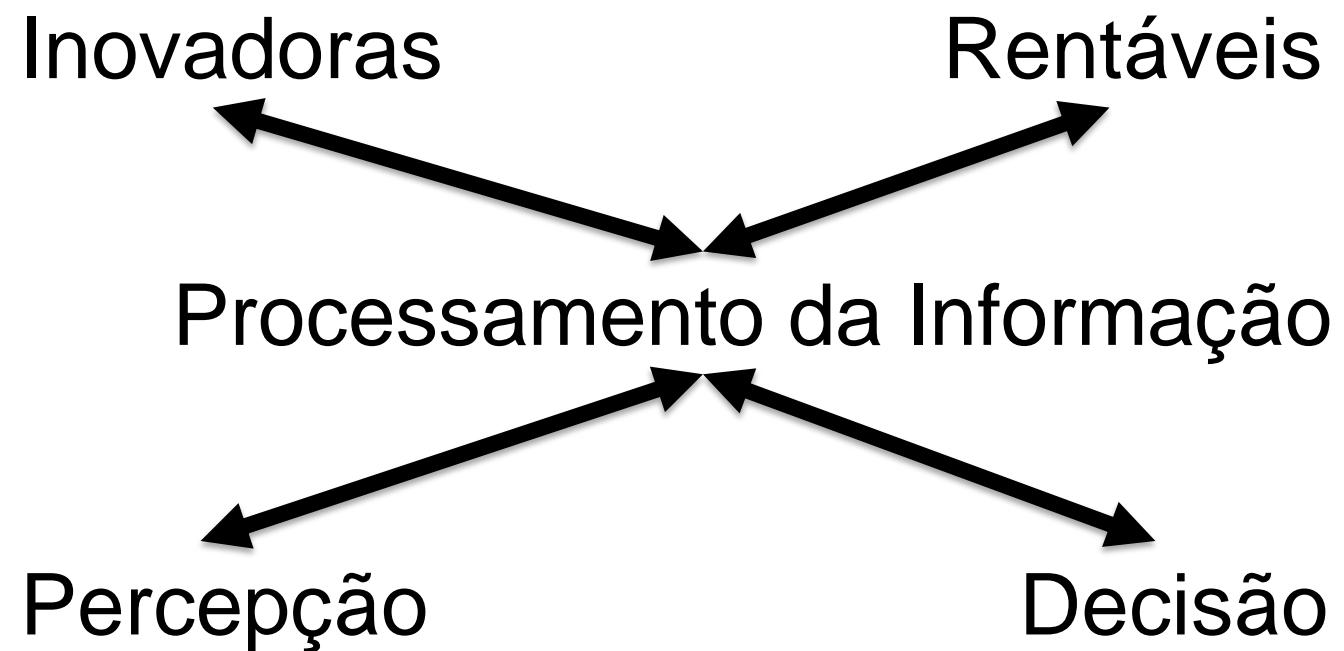


Introdução ao Big Data

- “Big Data faz referência ao grande **volume, variedade e velocidade** de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.” (*Gartner*)

Introdução ao Big Data

- “Big Data faz referência ao grande **volume, variedade e velocidade** de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.” (*Gartner*)



```
0110010101110010■0  
0000011000010111000  
1010011000010111000  
011011100110011100  
1110010■0111011101  
0100000011100000110  
1000001101001011011  
0110000101110000011
```

DATA IS THE NEW OIL



Data Monetization

Capitalizing on Your Data Assets



Data Driven Decisions



Introdução ao Big Data

- “Big Data faz referência ao grande **volume, variedade e velocidade** de dados que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.” (*Gartner*)

Os 3 Vs de Big Data

- **Os 3 Vs de Big Data**
 - Volume
 - Variedade
 - Velocidade



BIG DATA



VOLUME

DATA SIZE



VELOCITY

SPEED OF CHANGE



VARIETY

DIFFERENT FORMS
OF DATA SOURCES



VERACITY

UNCERTAINTY OF
DATA

Os 3 Vs de Big Data

- **Volume**

- Cerca de 2,5 bilhões de gigabytes de dados são criados diariamente;
- De toda a quantidade de dados disponível no mundo, aproximadamente 90% foi criado nos últimos 2 anos;
- 1,8 bilhão de usuários ativos no Facebook;
- 1 bilhão de usuários ativos no WhatsApp;
- 95 milhões de fotos e vídeos por dia no Instagram;
- 44 milhões de artigos na Wikipedia;
- 4 bilhões de visualizações por dia no YouTube;
- 300 horas de vídeos são carregados a cada 1 minuto no Youtube;

Os 3 Vs de Big Data

- **Velocidade**

- Processamento em tempo real e Streaming Data (dados em streaming).
 - O que acontece em 30 segundos na Internet.



Os 3 Vs de Big Data

- **Variedade**

- Dados são criados em diferentes formatos - como e-mails, comentários no Facebook, fotos publicadas em redes sociais e transações.
- Big Data inclui dados estruturados, semi-estruturados e não-estruturados.
 - Dados estruturados:
 - Bases de dados relacionais.
 - Dados semi-estruturados:
 - Não possuem uma estrutura pré-definida (representação estrutural heterogênea).
 - Auto-descritivos e sem esquema prévio definido.
 - Possuem esquema de representação presente (de forma explícita ou implícita).
 - Exemplo: XML (eXtensible Markup Language).
 - Dados não-estruturados:
 - Documentos, fotos, vídeos, *tweets*, comentários em redes sociais, etc.
 - Estima-se que pelo menos **80%** dos dados gerados atualmente sejam do tipo não-estruturados.

Outros Vs do Big Data

- Além dos 3 Vs, outras duas dimensões são comumente associadas à definição de Big Data. São elas:
 - **Veracidade**
 - Confiabilidade dos dados.
 - **Valor**
 - Gerar valor para o negócio;
 - Melhor entender as necessidades dos clientes;
 - Oferecer produtos e serviços que melhor atendam as necessidades dos clientes;
 - Oferecer produtos e serviços personalizados;
 - Melhorar o relacionamento com os clientes;
 - Aumentar a fidelização e satisfação dos clientes;
 - Gerar vantagem competitiva para o negócio.

Desafios do Big Data

- Onde armazenar esses dados?
- Como estruturar esses dados?
- Como consultar esses dados?
- Como extrair valor desses dados?
- Necessidade de novas tecnologias capazes de oferecer escalabilidade, disponibilidade, flexibilidade e desempenho para a manipulação de grandes volumes de dados.

Desafios do Big Data

- Big Data necessita de grande capacidade de processamento e armazenamento.
- Computação em Nuvem oferece capacidade de processamento e armazenamento conforme a necessidade do usuário.
- Computação em Nuvem é um imperativo para Big Data.
- De acordo com NIST (*National Intitute of Standards and Technology*): Computação em Nuvem é um modelo que permite um acesso sob demanda via redes de computadores a um conjunto compartilhado de recursos computacionais que podem ser rapidamente provisionado e liberado com um mínimo de esforço administrativo ou interação com o provedor de serviços.

Armazenamento

- SGBDs mais utilizados:

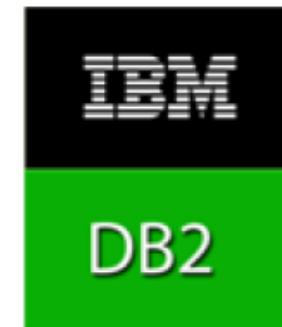


PostgreSQL



ORACLE®

TERADATA®



Armazenamento

- **Base de dados não relacionais**
 - **NoSQL (Not Only SQL)**
 - Conjunto de conceitos que permite o processamento rápido e eficiente de conjuntos de dados com foco em desempenho, confiabilidade e agilidade.
 - Diferentes formas de armazenamento
 - Orientado a documentos (o mais popular)
 - Ex.: MongoDB (<https://www.mongodb.com/>) e Apache CouchDB (<https://couchdb.apache.org/>)
 - Orientado a chave-valor (o mais simples)
 - Ex.: Amazon DynamoDB (<https://aws.amazon.com/dynamodb/>)
 - Orientado a grafos (o mais especializado)
 - Ex.: Neo4j (<https://neo4j.com/>)
 - Orientado a colunas (o mais complexo)
 - Ex.: HBase (<http://hbase.apache.org/>)
 - Características
 - Não-relacional
 - *Cluster-friendly*
 - Interface de consulta simples.

Armazenamento

- **Data Lake**

- Repositório único no qual dados estruturados e não-estruturados, coletados de diferentes fontes, são armazenados em sua forma bruta, como foram coletadas na fonte, sem qualquer processamento.

- **Enterprise Data Hub (EDH)**

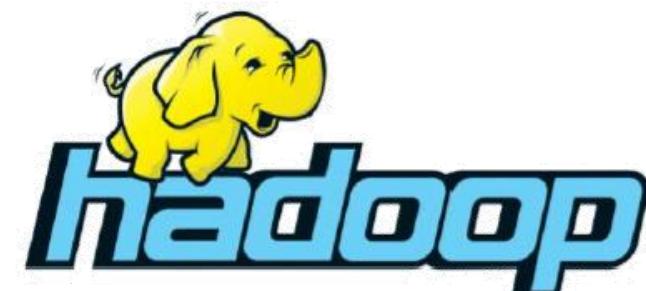
- Permite que a empresa tenha uma fonte de dados centralizada e unificada que possa fornecer rapidamente informações a diversos usuários do negócio, apoiando a tomada de decisão.

- Soluções:

- Azure Data Lake
 - <https://azure.microsoft.com/pt-br/solutions/data-lake/>
- Cloudera Enterprise Data Hub
 - <https://www.cloudera.com/products/enterprise-data-hub.html>
- Enterprise Data Hub (MapR)
 - <https://mapr.com/solutions/enterprise/enterprise-data-hub/>

Processamento

- **Hadoop**
 - <http://hadoop.apache.org/>
 - Solução *open source* que permite a execução de aplicações de Big Data utilizando milhares de máquinas.
 - Projetado para processar grandes quantidades de dados estruturados e não-estruturados.
 - Oferece recursos de armazenamento, gerenciamento e processamento de dados distribuídos.
 - Benefícios:
 - Redução de custo
 - Flexibilidade
 - Escalabilidade
 - Desempenho



Processamento

- Principais fornecedores de mercado

cloudera

MAPR[®]



Hortonworks

Processamento

- **Ecossistema Hadoop**
 - Hadoop possui 2 componentes principais:



Processamento

- **Hadoop HDFS (Hadoop Distributed File System)**



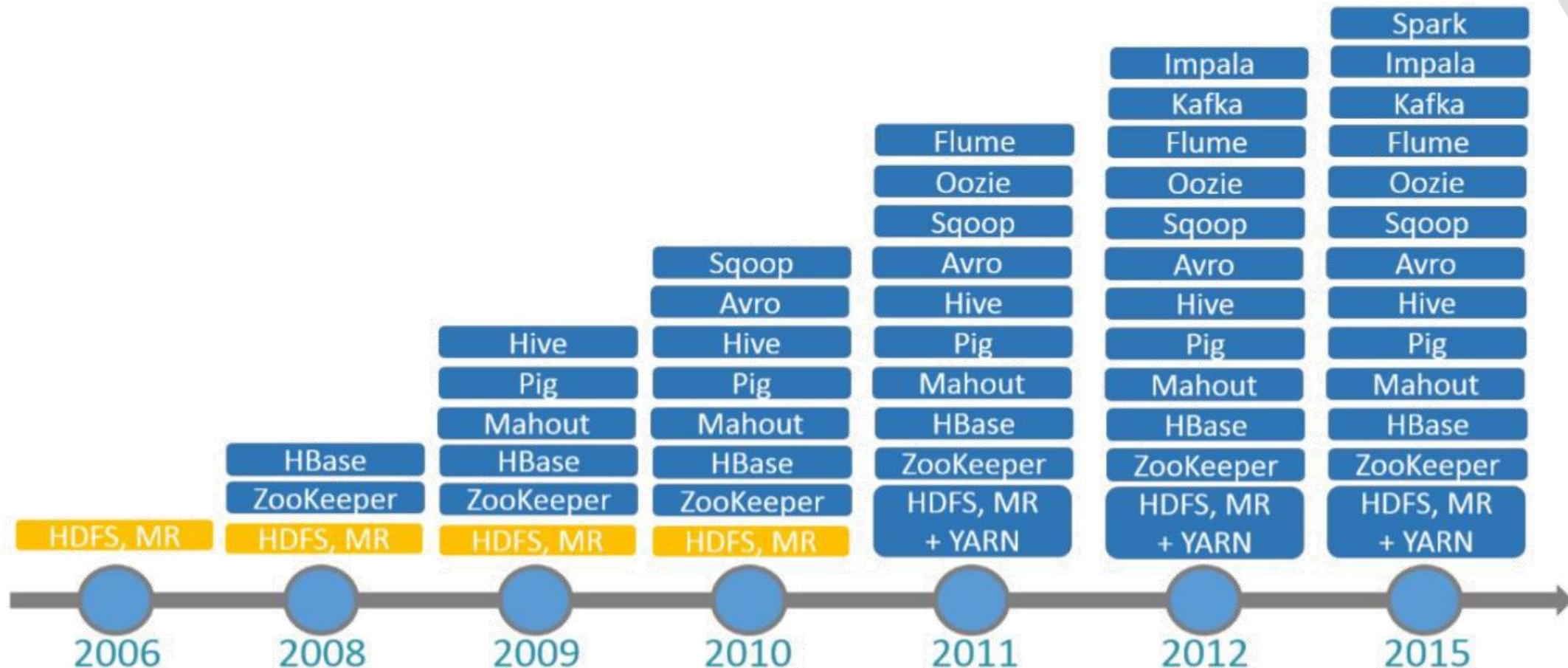
- Sistema de arquivos distribuídos;
- Otimizado para processamento de **grandes volumes de dados (alta taxa de transferência)**;
- Otimizado para ler e armazenar grandes arquivos em *clusters*;
- Arquivos são divididos em blocos de 64 ou 128 MB (tamanho *default* – *pode ser configurado*);
- Escalável e tolerante a falhas.

- **Hadoop MapReduce**



- É uma implementação do Hadoop;
- Ferramenta utilizada para facilitar o processamento de grandes volumes de dados (multi-terabyte data-sets) de forma distribuída;
- Tolerante a falhas;
- Funções *Map* e *Reduce*.

Ecossistema Hadoop



Processamento

- Algumas empresas que utilizam o Hadoop



MONSANTO



Aplicações

- **Varejo**
 - Melhor segmentação de clientes;
 - Propaganda personalizada;
 - Melhor oferta de produtos e serviços com base no perfil, rastro digital e no histórico de compras do cliente;
 - Previsão e prevenção de *Customer Churn* (identificação de clientes com alta propensão a cancelar produtos e serviços);
 - *Chatbots*.
- **Setor financeiro**
 - Detecção de transações fraudulentas envolvendo utilização de Internet Banking e cartões de crédito;
 - Análise de crédito;
 - Melhor relacionamento com os clientes.

Aplicações

- **People Analytics (HR Analytics)**

- Processo de coleta, armazenamento e análise de dados sobre o comportamento dos colaboradores em uma organização.
- Utilização de análise de dados em Gestão de Pessoas.
- Utilização de informações disponíveis em redes sociais.
- Utilização de leitores biométricos e crachás inteligentes.
- Análise de currículo utilizando *Text Analytics* (Análise de Texto).
- Utilização de rastro digital e informações de redes sociais para ajudar na identificação do perfil mais adequado para cada vaga.
- Principais benefícios:
 - Otimização do processo de Recrutamento e Seleção;
 - Avaliação de Desempenho;
 - Aumento da produtividade;
 - Desenvolvimento de programas de treinamento e capacitação;
 - Retenção de talentos;
 - Redução da Rotatividade.

Aplicações

- **Internet of Things (Internet das Coisas)**

- Rede formada por milhares de dispositivos (objetos) inteligentes conectados a Internet.
- Dispositivos capazes de capturarem grandes quantidades de dados por meio de sensores.
- Exemplos:
 - Veículos autônomos e conectados;
 - Casas inteligentes;
 - Eletrodomésticos inteligentes;
 - *Wearables* (Relógios e pulseiras inteligentes).

Exemplos de empresas com negócios centrados em dados



U B E R



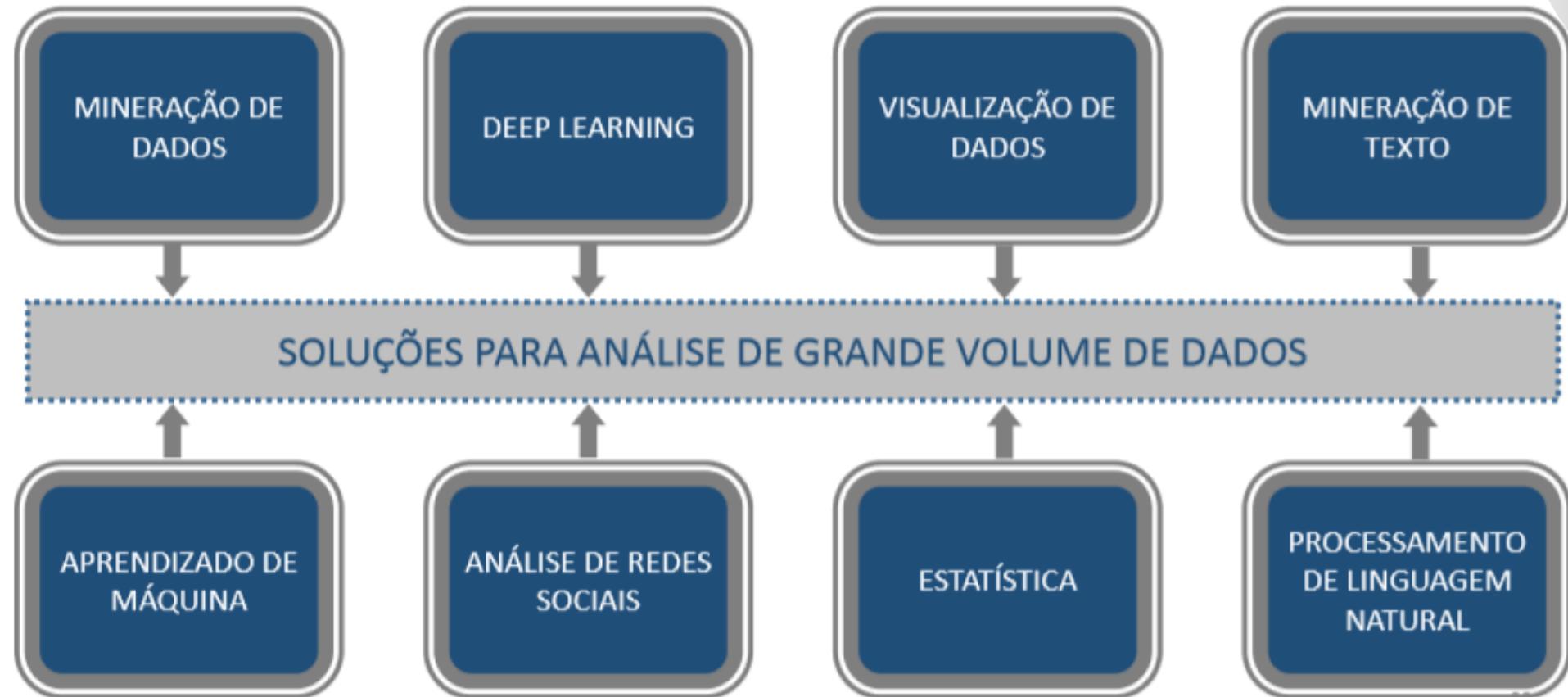
Parte II – Definindo Data Science

Introdução a Data Science

- **Data Science (Ciência dos Dados)**

- Termo utilizado para descrever o processo de extração, análise e interpretação de grandes volumes de dados, gerados à partir de diversas fontes, a fim de extrair insights e informações valiosas para auxiliar na tomada de decisões.
- Incorpora conhecimentos das áreas de Matemática e Estatística e utiliza diversas técnicas, como Modelagem Preditiva, Mineração de Dados (*Data Mining*), Análise de Texto (*Text Analysis*), Aprendizado de Máquina (*Machine Learning*) e Visualização de Dados (*Data Visualization*).

Inovação na análise de dados



Técnicas utilizadas em Análise de Dados

- ***Data Mining (Mineração de Dados)***

- Processo que tem por objetivo analisar grandes quantidades de dados a fim de identificar padrões e extrair informações.
 - Exemplo: Weka (<https://www.cs.waikato.ac.nz/ml/weka/>)

- ***Machine Learning (Aprendizado de Máquina)***

- Subcampo da Inteligência Artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitem ao computador aprender à partir do reconhecimento de padrões.
 - Exemplo: IBM Watson Machine Learning (<https://www.ibm.com/cloud/machine-learning>).

- ***Deep Learning (Aprendizado Profundo)***

- Subárea de Maching Learning.
- Aprendizado de máquina utilizando redes neurais artificiais.
 - Processamento de Linguagem Natural;
 - Reconhecimento de Fala;
 - Visão Computacional;
 - Processamento de Imagens.

Exemplo IBM Watson Studio

Define your project goals

What do you want to
find out?

Do you have the data to
analyze?

Prepare the data

Refine the data
Add the data as a
project asset or in a
data repository

Choose a tool

Pick the tool that
matches your data and
desired outcome
Choose between an
automated process, a
graphical editor, or
code your own model

Train your model

Train the model with
the data you supply
Let a model building
tool choose estimators
and optimizers or
choose your own

Deploy your model

Score the model to
generate predictions
Make your model
available in production
Retrain as needed

Definir suas
metas

Prepare seus
dados

Escolha uma
ferramenta

Treine seu Modelo

Implemente seu
Modelo

Técnicas utilizadas em Análise de Dados

- **Estatística**
 - Conjunto de métodos largamente utilizados na coleta e interpretação de dados em *Data Science*.
- **Análise de Texto (*Text Analysis*)**
 - Análise de dados não estruturados (textos).
 - Utiliza Processamento de Linguagem Natural.
 - Subcampo da Inteligência Artificial que estuda a compreensão da linguagem natural (Análise Semântica e Análise sintática).
 - Técnica utilizada para realizar análise de conteúdo em mídias sociais.
- ***Data Storytelling***
 - Técnica de construção de narrativas por meio da utilização de visualização de dados.

Inovação na análise de dados

80%

do processo de análise é gasto **preparando os dados**



Oportunidades

- Big Data Analytics
- Serviços orientados a dados
- Monetização de dados
- Ferramentas para análise de dados
- Serviços de visualização de dados

Big Data Analytics

- Processo de coletar, organizar e analisar enormes conjuntos de dados a fim de se descobrir padrões, tendências e fazer correlações.
- Faz uso de dados históricos, mineração de dados, aprendizado de máquina, entre outros métodos.
- Quatro abordagens:
 - Descritiva
 - Diagnóstica
 - Preditiva
 - Prescritiva

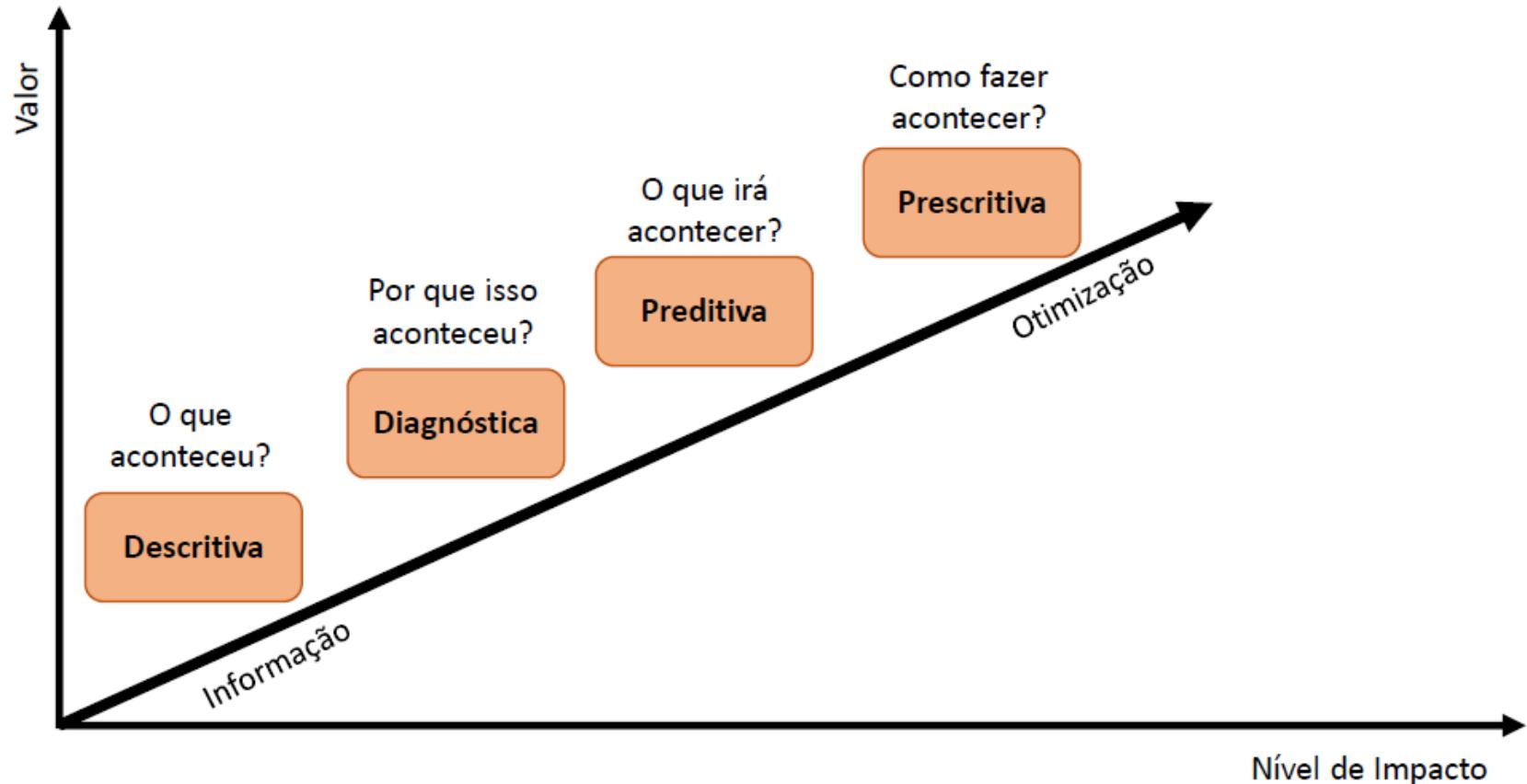
Big Data Analytics

- **Descritiva**
 - Entender os eventos ocorridos.
 - Faz uso de ferramentas de BI (*Business Intelligence*);
 - Baseado em métricas;
 - *Dashboards*, relatórios e alertas;
 - Estima-se que mais de 80% das análises de negócios realizadas sejam descritivas.
- **Diagnóstica**
 - Tem por objetivo determinar a causa de um evento;
 - Uso de análise de correlação, análise de variância e testes de hipótese.

Big Data Analytics

- **Preditiva**
 - Prever tendências baseadas em dados;
 - Predição de eventos futuros;
 - Faz uso de diferentes métodos e ferramentas (análises estatísticas, técnicas de simulação, mineração de dados e aprendizado de máquina);
 - Utilizada para diferentes aplicações, como por exemplo, prever quais clientes de uma operadora de telefonia estariam mais propensos a cancelarem um determinando serviço.
- **Prescritiva**
 - Predizer as possíveis consequências para as diferentes escolhas que forem feitas;
 - Sugere ações baseadas no conhecimento extraído dos dados.

Big Data Analytics



Serviços orientados a dados

- **Web sites**
 - Análise de experiência do usuário
 - Personalização de conteúdo
- **Geolocalização**
 - Recomendação de serviços
- **Comércio eletrônico**
 - Recomendação de produtos
 - Segmentação de clientes
- **E-mails e mensagens**
 - Publicidade personalizada
- **Redes sociais**
 - Análise de influência
 - Análise de sentimento

Monetização de dados

- Transformar ativos de informação em dinheiro, direta ou indiretamente, por meio de troca, comercialização ou venda direta.
 - Dados genéticos para pesquisadores;
 - Dados de comportamento de usuários para campanhas de marketing;
 - Dados de sensores para seguradoras.

Ferramentas para análise de dados

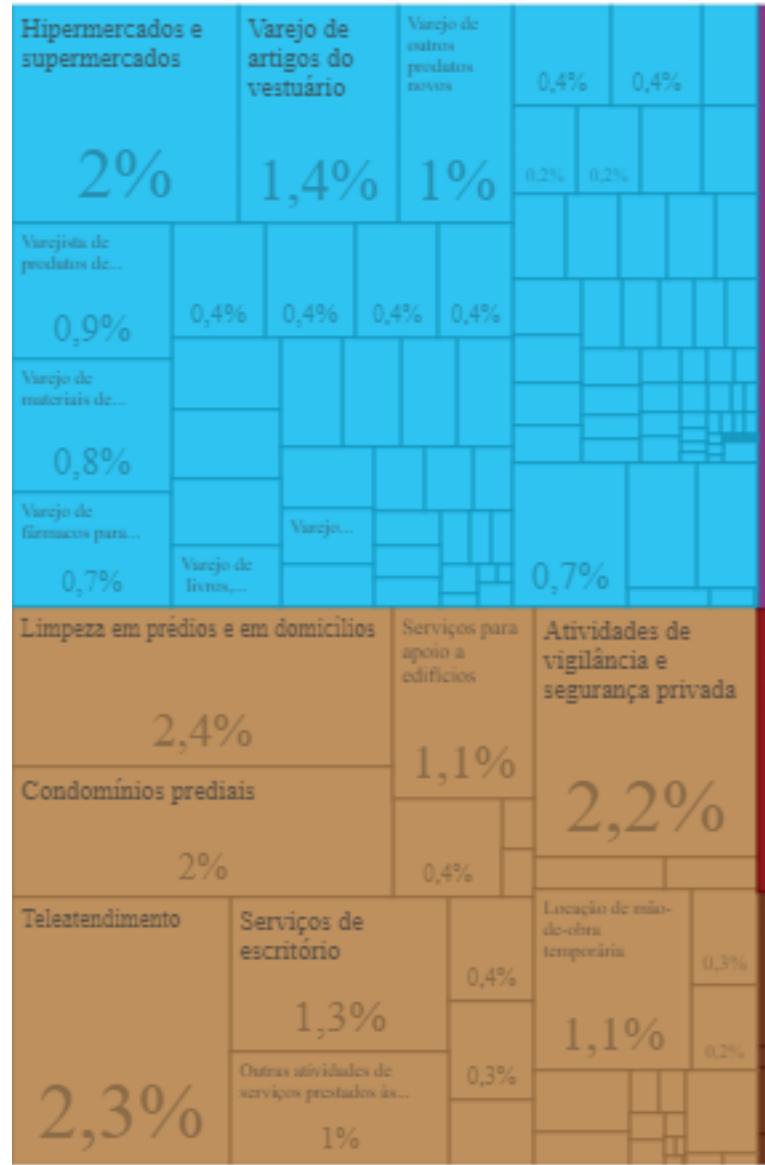
- Processamento distribuído dos dados;
- Processamento em tempo real;
- Algoritmos de aprendizado de máquina (*Machine Learning*).

Serviços de visualização de dados

- Representação gráfica de informações (utilização de gráficos de barra, gráficos de linha, gráficos de área, imagens, mapas, entre outros).
- Tem por objetivo facilitar a compreensão dos dados.
- **Benefícios**
 - Tomada de decisão aperfeiçoada;
 - Monitoramento e aumento da produtividade;
 - Melhoria na análise de dados;
 - Melhor experiência ao usuário.
- Exemplos:
 - DataViva – Portal de visualização de dados da economia brasileira. Disponibiliza dados socioeconômicos de mais de 5 mil municípios brasileiros.
 - <http://dataviva.info/pt/>
 - <http://circos.ca/>
 - Tweetping - <https://tweetping.net/>
 - Infogr.am <https://infogr.am/>

Indústrias em São Paulo (2017)

Empregos: 4,87 M

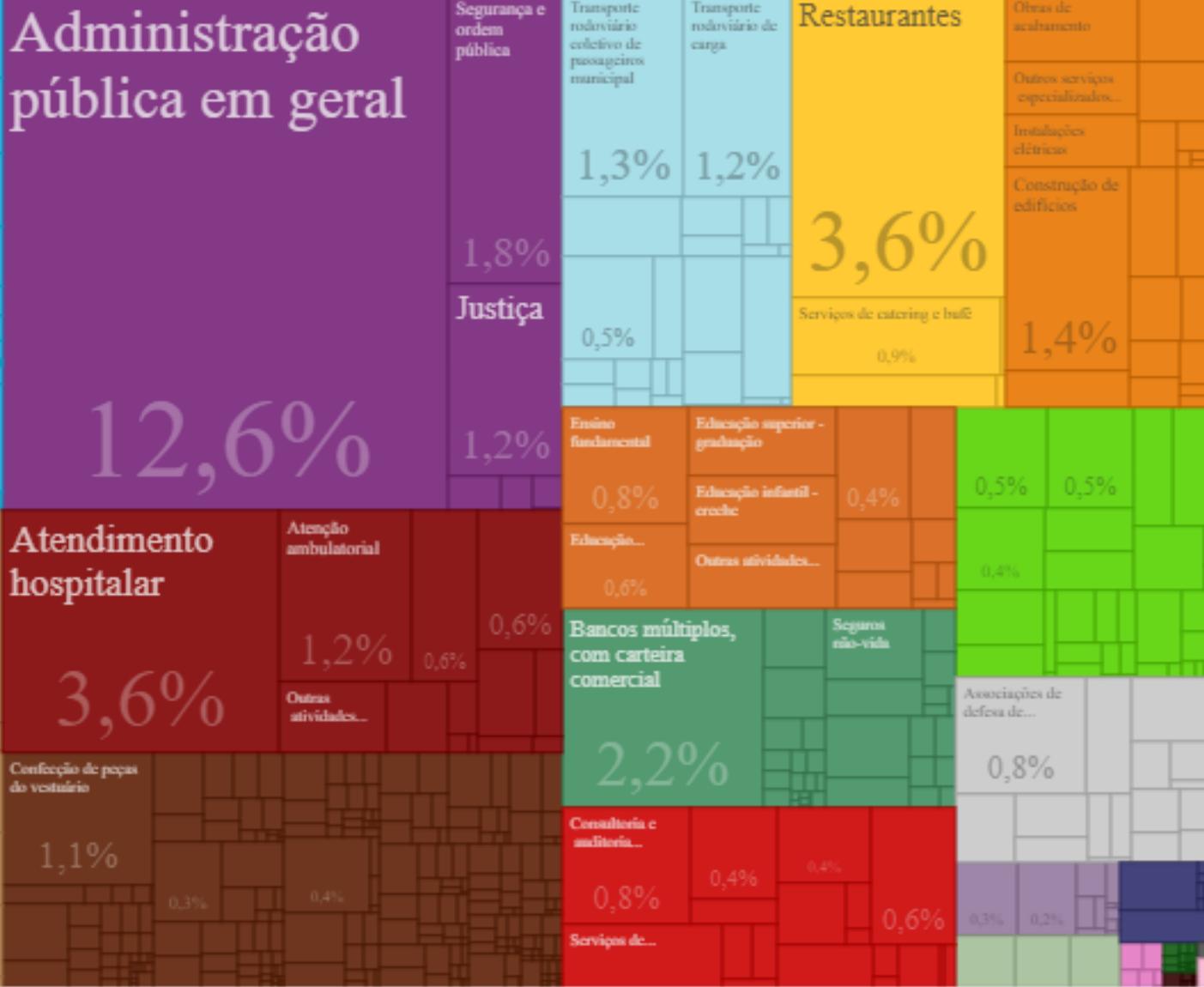


Administração pública em geral

12,6%

Atendimento hospitalar

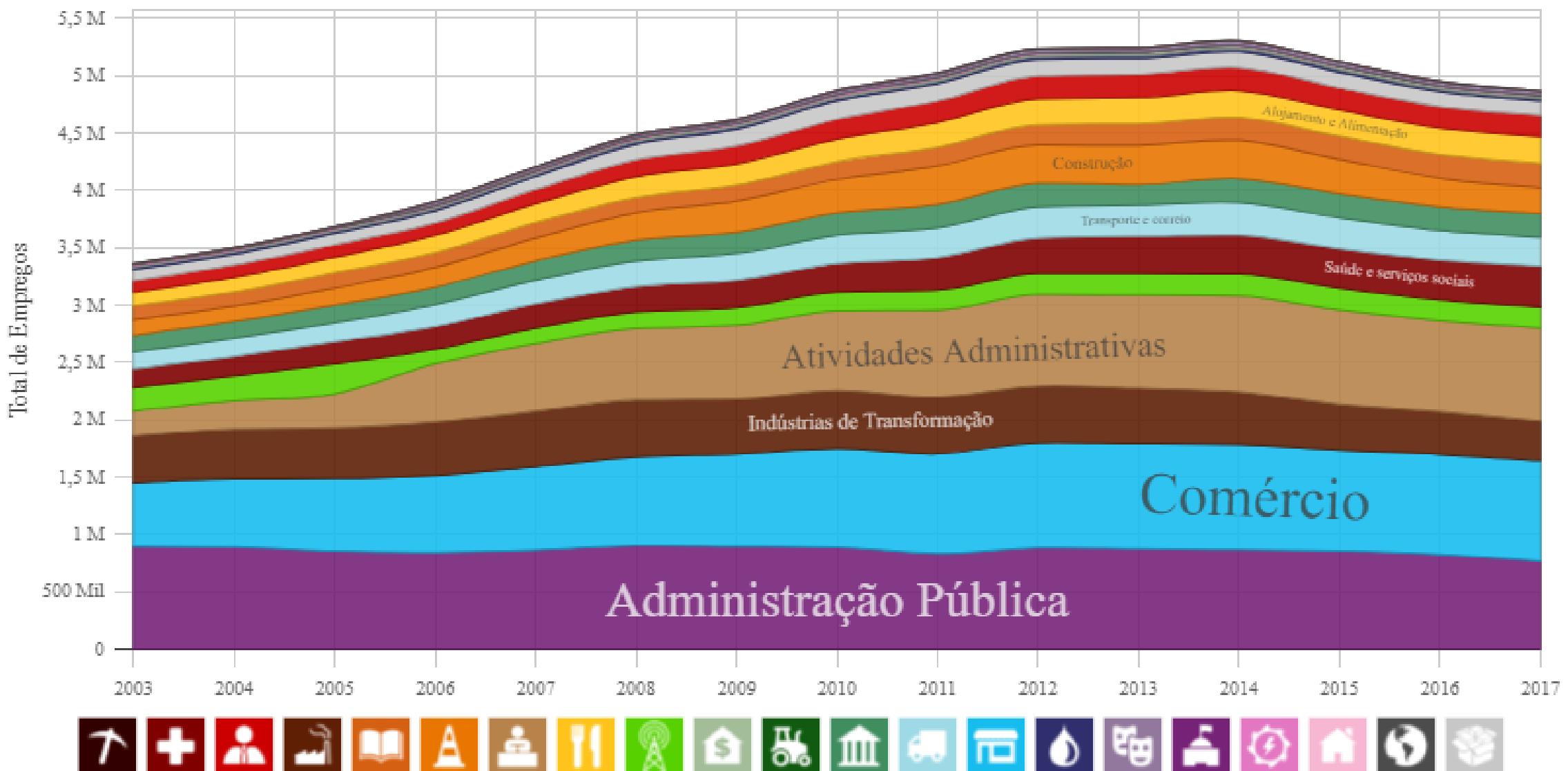
3,6%



Dados fornecidos por RAIS

Indústrias em São Paulo (2003-2017)

Empregos: 68,4 M



Dados fornecidos por RAIS

Empregos por Família (2017)

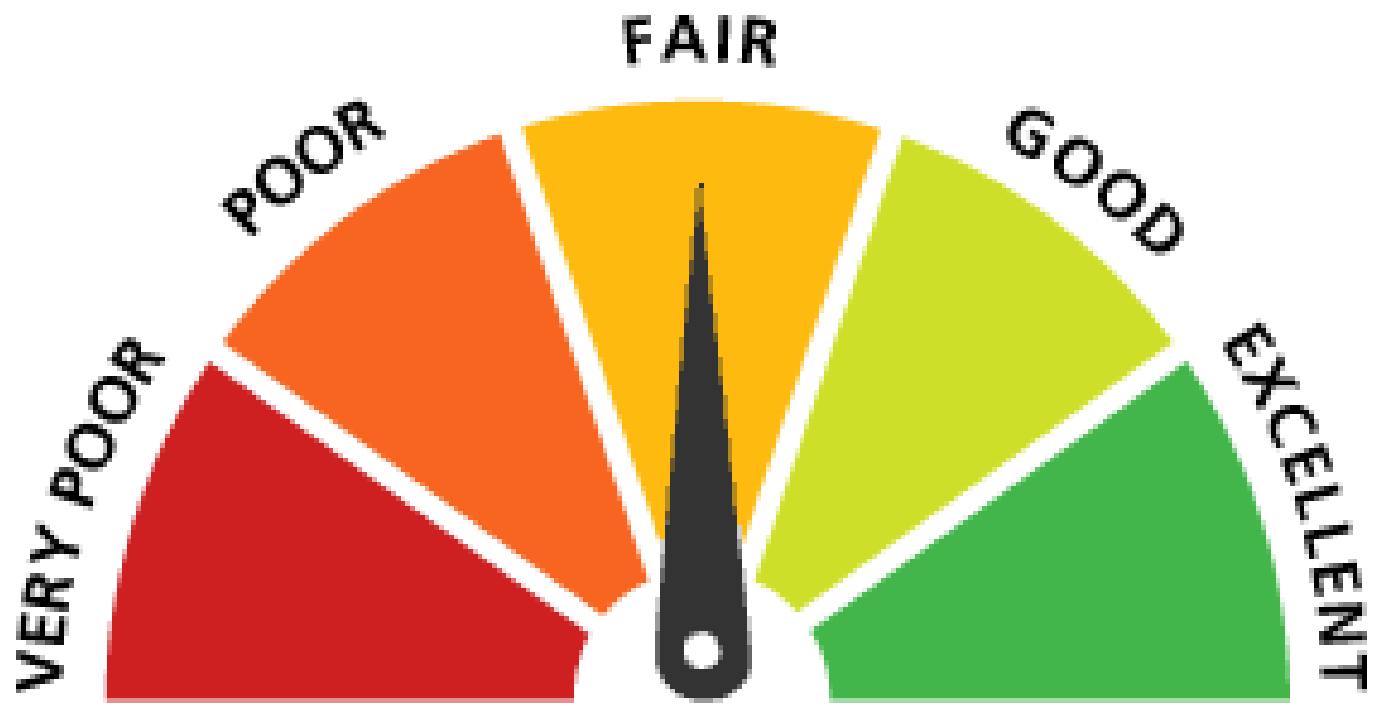
Empregos: 4,77 M



Dados fornecidos por RAIS



**Use case:
credit scoring**



Fluxo na plataforma



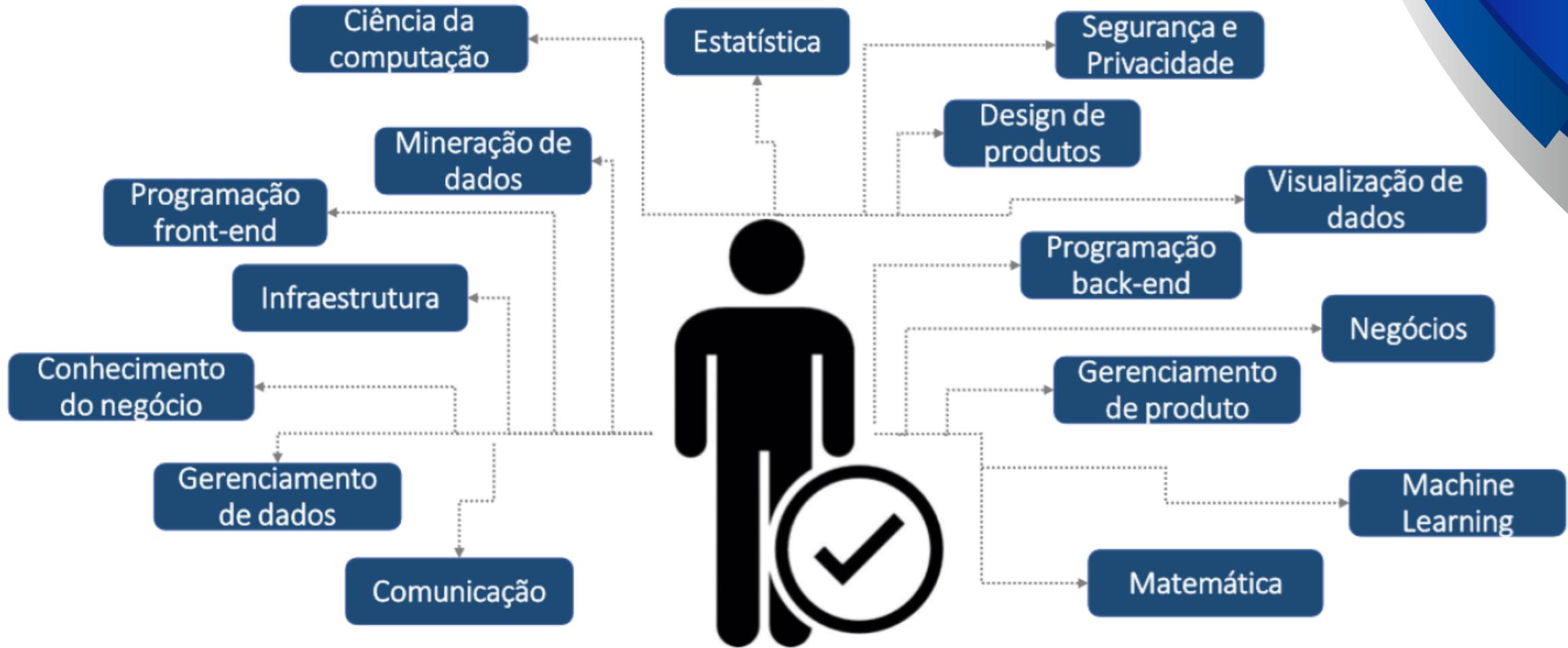
Linguagens de programação utilizadas em Data Science

- **Linguagem R**
 - Linguagem de programação estatística largamente utilizada em Big Data.
 - <https://www.r-project.org/>
- **Python**
 - Python é uma das linguagens de programação orientada a objetos mais versáteis, fáceis e rápidas de serem aprendidas.
 - <https://www.python.org/>
- **Scala (Scalable Language)**
 - Scala é uma linguagem de programação que combina os paradigmas de programação orientada a objetos e funcional;
 - Utiliza uma sintaxe concisa que é totalmente compatível com Java e é executado na JVM (*Java Virtual Machine*);
 - Tem ganhado cada vez mais adeptos por conta da sua capacidades de lidar com grandes quantidades de dados de maneira escalável e confiável;
 - <http://www.scala-lang.org/>

Cientista de dados

- Profissional responsável por extrair informações de grandes volumes de dados estruturados e não estruturados;
- Profissão em alta;
- Salários podem chegar a 22 mil reais (Robert Half/Computerworld);
- Carência de profissionais qualificados;
- Profissionais capacitados são altamente disputados;
- Carreira altamente promissora.

Cientista de Dados



Introdução à Computação em Nuvem

- Computação em Nuvem é um modelo que provê acesso sob demanda via rede de computadores a um conjunto compartilhado de recursos computacionais que pode ser rapidamente provisionado e liberado com um mínimo de esforço administrativo ou interação com o provedor de serviços (NIST - *National Institute of Standards and Technology*).

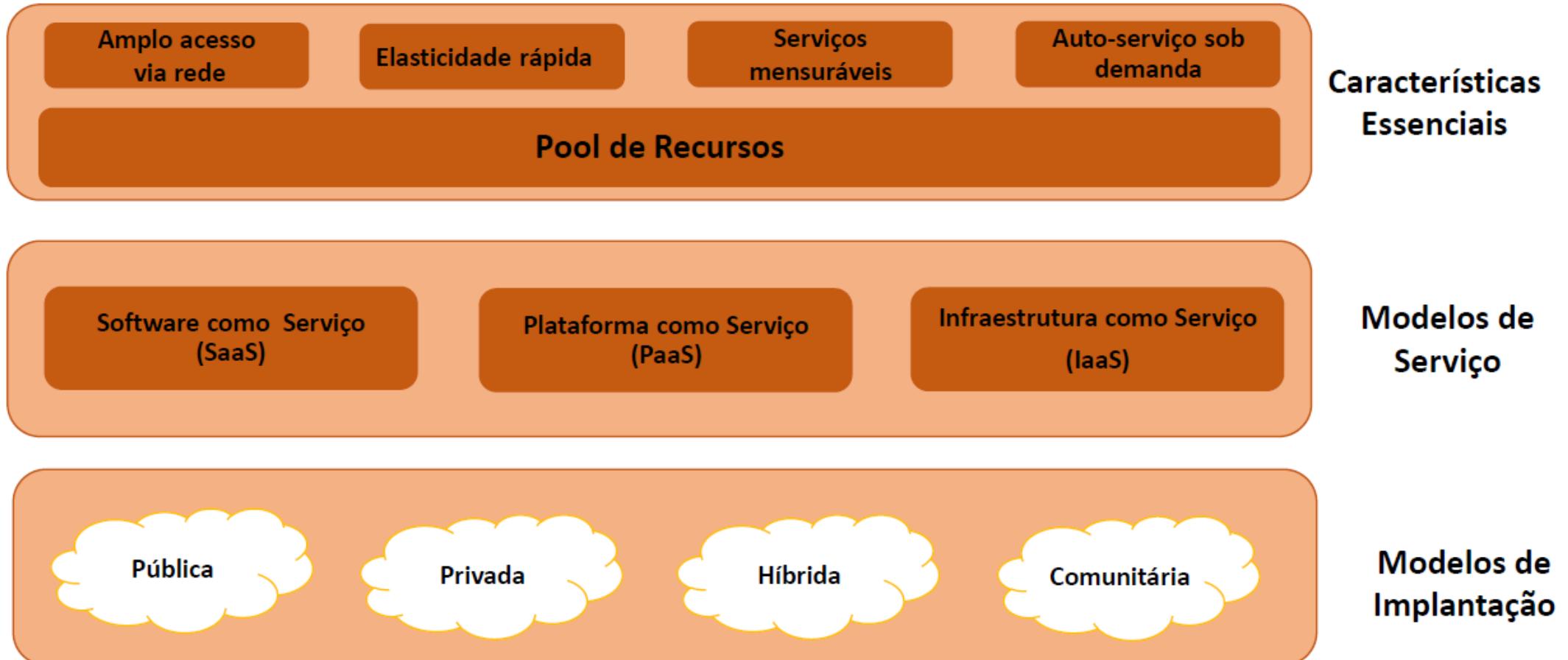
Características essenciais de acordo com o NIST

- **Auto-serviço sob demanda**
 - Usuário pode provisionar recursos computacionais conforme necessidade, sem demandar interação manual do provedor, podendo ser de forma automática ou não.
- **Amplo acesso via rede**
 - Recursos devem estar disponíveis via rede (tipicamente via Internet).
 - Acesso via múltiplas plataformas.
- **Pool de recursos**
 - Recursos físicos e virtuais dinamicamente alocados de acordo com a demanda do usuário.
Exemplo: armazenamento, processamento, memória e banda.
- **Elasticidade rápida**
 - Recursos podem ser rapidamente provisionados para atender o aumento da demanda. De forma análoga, recursos podem ser rapidamente desalocados caso não haja demanda.
- **Serviços mensuráveis**
 - Métricas de uso e tarifação.

Modelos de implantação

- **Nuvem pública**
 - Provedor fornece os recursos de computação, como servidores e armazenamento pela Internet.
- **Nuvem privada**
 - Modelo no qual os recursos são utilizados exclusivamente por uma única empresa ou organização.
- **Nuvem híbrida**
 - Modelo que combina nuvens públicas e privadas.
- **Nuvem comunitária**
 - Modelo no qual a infraestrutura da nuvem é compartilhada por grupos de organizações com interesses em comum.

Modelo visual da definição de Computação em Nuvem do NIST



Modelos de serviço

- **Infraestrutura como um serviço (*Infrastructure as a Service* - IaaS)**
 - Fornece infraestrutura computacional (servidores, máquinas virtuais, armazenamento, redes e sistemas operacionais) como serviço de forma provisionada e gerenciada pela Internet.
 - Amazon EC2, Microsoft Azure, etc.
- **Plataforma como um serviço (*Platform as a Service* - PaaS)**
 - Fornece um ambiente sob demanda para desenvolvimento, teste e gerenciamento de aplicativos de software e que permite aos desenvolvedores criarem aplicativos, sem se preocupar com a configuração ou o gerenciamento de infraestrutura de servidores, armazenamento, rede e bancos de dados necessários para desenvolvimento.
 - Salesforce.com
- **Software como um serviço (*Software as a Service* - SaaS)**
 - Permite às empresas fornecer aplicativos de software pela Internet, sob demanda, geralmente, em forma de assinaturas:
 - Gmail, Flickr, Google Apps, etc.

Provedores de Computação em Nuvem

- **Amazon Elastic Compute Cloud**
 - Serviço de Computação em Nuvem da Amazon.
 - <https://aws.amazon.com/>
- **Microsoft Azure**
 - Plataforma de Computação em Nuvem da Microsoft.
 - <https://azure.microsoft.com/>
- **Google Cloud Platform**
 - Plataforma de Computação em Nuvem do Google.
 - <https://cloud.google.com/>
- **OpenStack**
 - Software *open-source* largamente utilizado para criação de nuvens privadas.
 - Fundado pela Rackspace e NASA.
 - <https://www.openstack.org/>

Apache Spark

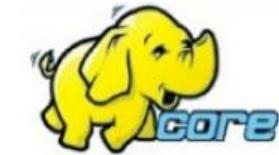


- Apache Spark é um poderoso mecanismo de processamento de código aberto construído em torno de velocidade, facilidade de utilização, e análises sofisticadas.
- Ela foi originalmente desenvolvida na Universidade de Berkeley em 2009.
- Desde o seu lançamento, Spark tem visto uma rápida adoção por parte das empresas em uma ampla gama de indústrias.
- Potências da Internet como Netflix, Yahoo e eBay implantou Spark em escala maciça, processar coletivamente múltiplos petabytes de dados em clusters de mais de 8.000 nós.
- Ele rapidamente se tornou a maior comunidade open source em big data, com mais de 1000 colaboradores de mais de 250 organizações.



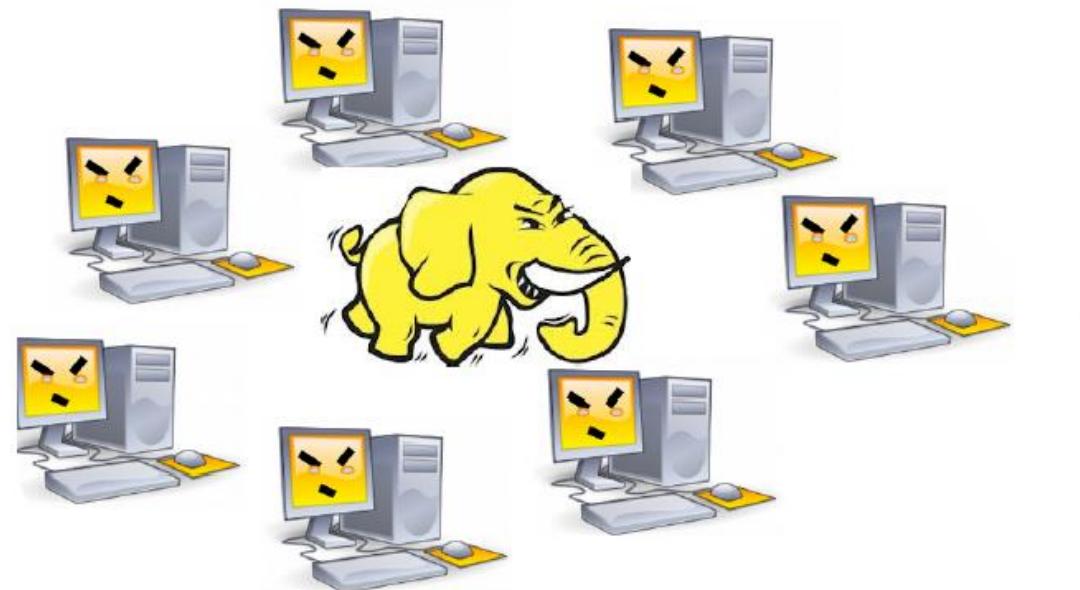
- O Apache Hadoop é um projeto de software open-source escrito em Java. Escalável, confiável e com processamento distribuído.
- Filesystem Distribuído.
- Inspirado Originalmente pelo GFS e MapReduce da Google (Modelo de programação MapReduce)
- Utiliza-se de Hardware Comum (Commodity cluster computing)
- Framework para computação distribuída infraestrutura confiável capaz de lidar com falhas (hardware, software, rede)

Ecosistema - Hadoop



MapReduce - Programação Distribuída

- modelo de programação para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes.



MongoDB

- Banco de dados não relacional (NoSQL) Orientado a Documentos
- Baseado em JSON onde os documentos (registros) são representados por “chave:valor” BSON
- Escrito em C++ e Open Source
- Schema Dinâmico: Permite dados complexos não estruturados
- Documentos auto-contidos e arrays reduzem a necessidade de join's
- Multiplataforma e com Alta Performance



Apache Cassandra

- É um tipo de banco NoSQL que originalmente foi criado pelo Facebook e atualmente é mantido pela Apache e outras empresas.
- Banco de dados distribuído baseado no modelo BigTable do Google e no Dynamo da Amazon



Data Analysis & Platforms



Databases / Data warehousing



Operational



Multivalue database



Business Intelligence



Data Mining



Social



KeyValue



Document Store



Graphs



Object databases



Multimodel



XML Databases



Parte III Big Data em sistemas IoT

Introdução

- Big Data em IoT é uma área grande e de rápido desenvolvimento onde muitos métodos e técnicas diferentes podem desempenhar um papel.
- Devido ao rápido progresso no aprendizado de máquina e novos desenvolvimentos de hardware, uma mudança dinâmica de métodos e tecnologias pode ser observada.
- Esta visão geral, portanto, tenta ser ampla e de alto nível, sem pretender ser abrangente. Sua abordagem em relação a Big Data e IoT é baseada em uma distinção entre a economia digital e as características do que Robin Milner descreveu como o Sistema de Computação Ubíquo (UCS) (Milner, 2009).

Sistemas de computação ubíqua - Ubiquitous Computing Systems (UCS)

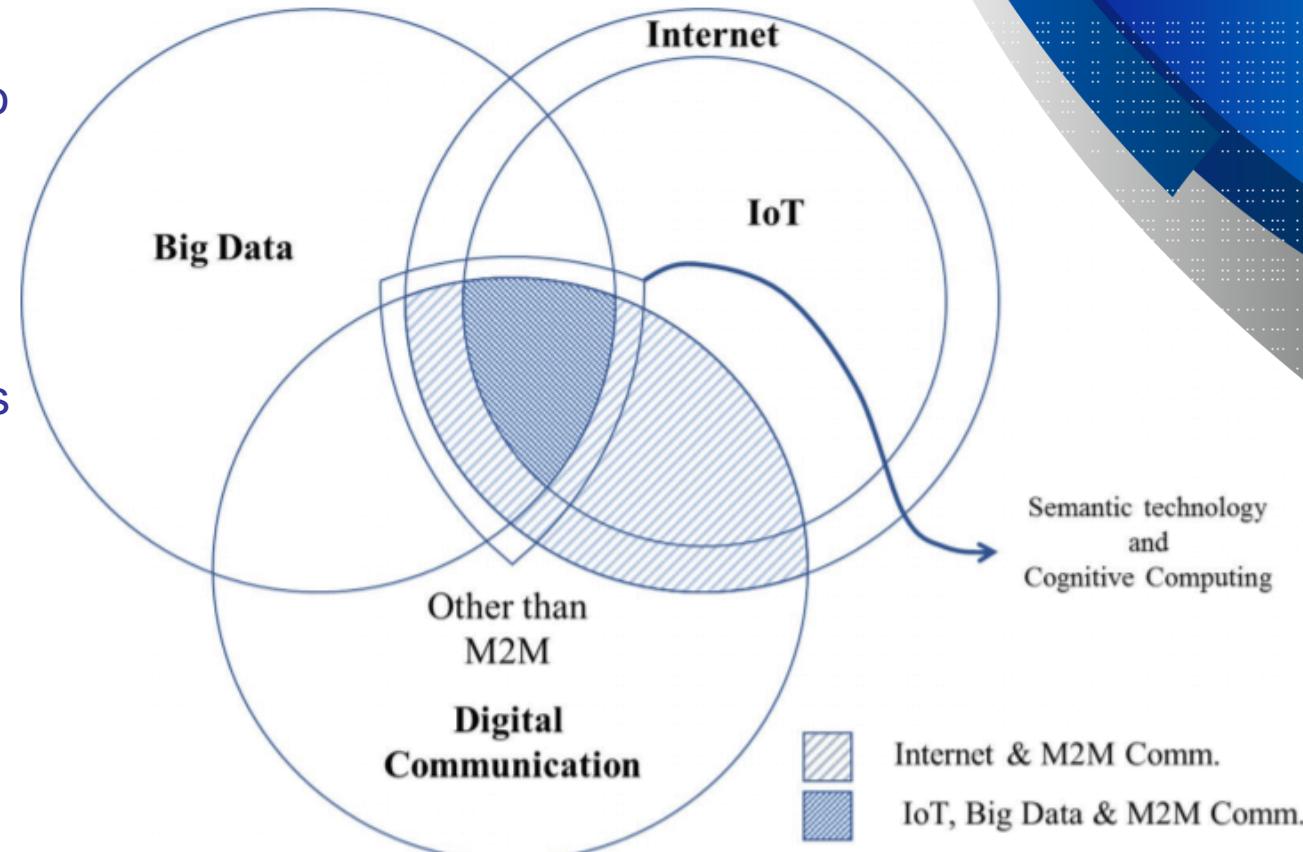
- As características de um UCS são (Milner, 2009):
 - Continuamente tomará decisões até então tomadas por nós;
 - Será vasto, talvez 100 vezes os sistemas de hoje; (
 - Deve se adaptar continuamente on-line aos novos requisitos; e
 - UCSs individuais irão interagir uns com os outros
- Milner (2009) define o UCS como:
 - um sistema com uma população de agentes interativos que gerenciam algum aspecto de nosso ambiente.
- Por sua vez, esses agentes de software se movem e interagem, não apenas no espaço físico, mas também no espaço virtual.
- Eles incluem estruturas de dados, mensagens e uma hierarquia estruturada de módulos de software.
- A visão formal de Milner é de uma torre de linguagens de processo que podem explicar a computação ubíqua em diferentes níveis de abstração.
- Os recursos genéricos do UCS contemporâneo incluem simultaneidade, interação e controle descentralizado.

Economia Digital

- A noção de economia digital foi claramente articulada na definição do programa Industry 4.0 da Alemanha onde o gerenciamento de manufatura e o setor de software convergem em um conceito conjunto que combina TI, análise de Big Data e produção em escala global.
- Enquanto as máquinas de manufatura industrial se comunicam dentro da IoT, os técnicos humanos devem ter a capacidade de verificar a produção e a qualidade do processo localmente em um local de produção e, eventualmente, tomar decisões em tempo real com base em análises complexas fornecidas pelos componentes de análise de dados da indústria global 4.0.
- O software de inteligente e baseado em nuvem (Gottwelles, 2016) permite que se integre o técnico local na IoT da Indústria 4.0.
- Os sistemas de software de fábrica digital global para otimização da Indústria 4.0 tornaram-se uma ferramenta de controle central desenvolvida pelos principais fabricantes e desenvolvedores de software como Siemens, Bosch, Kuka, SAP e Fraunhofer IPA.
- Embora o monitoramento em tempo real, a análise e a rastreabilidade constituam um conjunto de aspectos em que as técnicas de Big Data e o aprendizado de máquina entram em ação, outro aspecto é a previsão e a modelagem.
- Novamente, grandes sistemas de software com módulos analíticos de Big Data são necessários para executar simulações virtuais de complexos processos de produção, logística, distribuição, risco financeiro, saúde do equipamento e aspectos humanos, etc.

Big Data e IoT

- Para fornecer uma visão geral dos vários domínios da comunicação digital associados ao Big Data e aos sistemas IoT, o diagrama ao lado que descreve duas áreas de interseção.
- As linhas inclinadas claras marcam a interseção entre "Comunicação Digital" e "A Internet", enquanto as linhas inclinadas escuras identificam o domínio de interseção entre "Big Data", "Internet das Coisas" (IoT) e "Comunicação Máquina a Máquina".



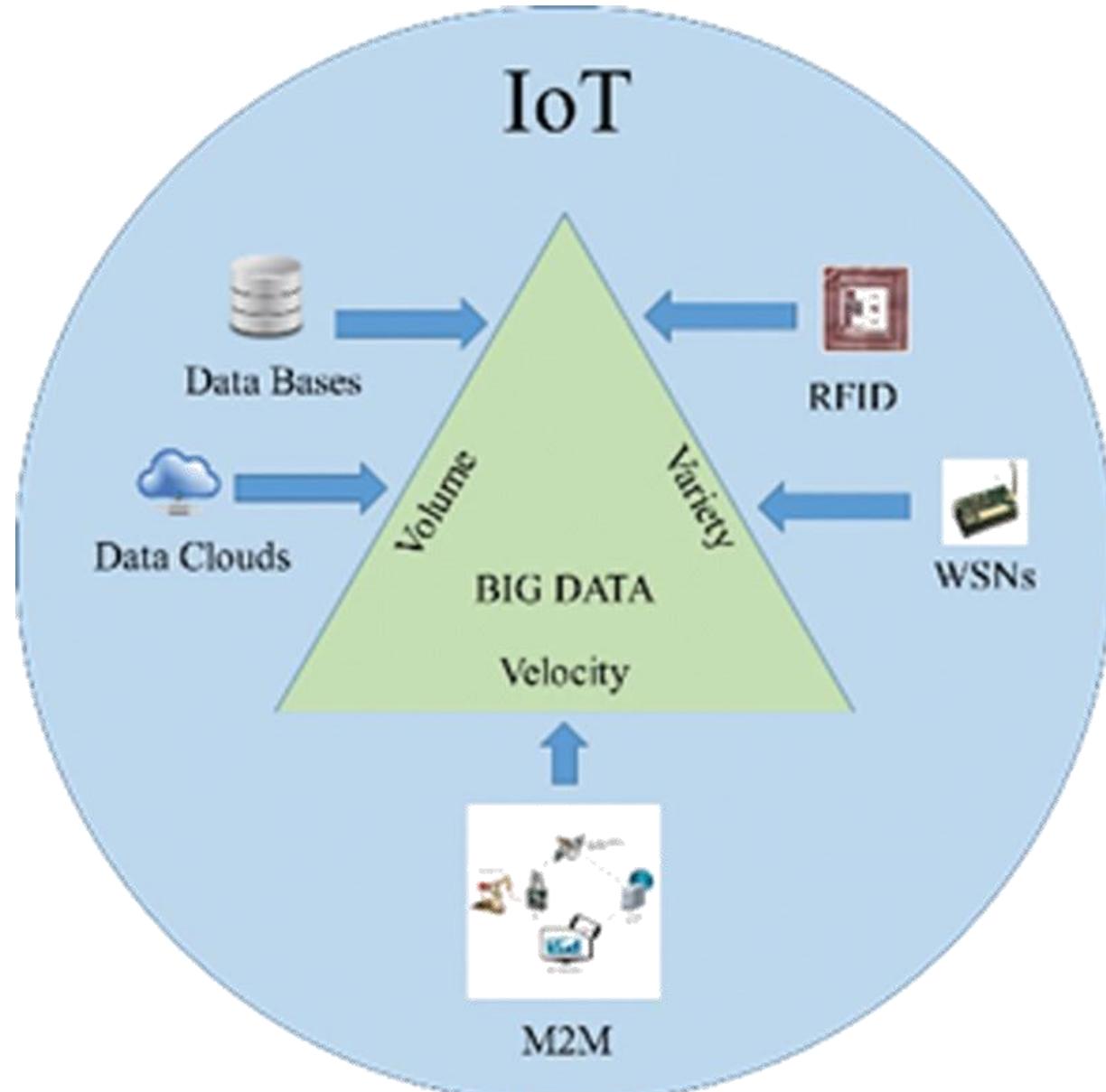
A Internet das Coisas

- A Internet das Coisas (IoT) é uma das plataformas emergentes mais rapidamente para a economia digital (Juniper, 2018).
- A IoT é uma rede gigante de “coisas” conectadas (que também inclui pessoas). Muitos países desenvolvidos estão aplicando ou planejando aplicar a IoT em casas e cidades inteligentes.
- O IoT European Research Cluster (IERC) propôs uma série de projetos de IoT e criou um fórum internacional de IoT para desenvolver uma visão estratégica e técnica conjunta para o uso de IoT na Europa (Santucci, 2010).
- A China planeja investir US \$ 166 bilhões nas indústrias de IoT até 2020 (Voigt, 2012).
- A crescente IoT produz uma grande quantidade de dados que precisarão ser processados e analisados.
- Embora não haja uma definição clara para Big Data, uma caracterização comumente citada são os “3Vs”: volume, variedade e velocidade (Laney, 2001; Zaslavsky et al.
- Foi previsto por vários autores que o grande número de objetos conectados na IoT gerará uma enorme quantidade de dados (Botta Zhang e outros)
- Os dados gerados pela IoT são variáveis em termos de estrutura, muitas vezes chegam em tempo real e podem ser de proveniência incerta.

Big Data e suas fontes

- Ao contrário da Internet convencional com a especificação padrão, atualmente a IoT não possui uma arquitetura de sistema tão bem definida (Huansheng Ning & Hu, 2012).
- Nos últimos cinco anos, diferentes tipos de estruturas foram propostos para a arquitetura do sistema IoT, incluindo modelos baseados em camadas, modelos baseados em dimensões, estruturas de domínio de aplicativo e estruturas de domínio social (Luigi Atzori et al., 2010; Huansheng Ning & Hu, 2012)
- Modelos baseados em camadas são as estruturas mais comumente usadas na literatura de IoT e camadas na estrutura são normalmente camadas de sensor, camadas de rede, camadas de serviço e camadas de interface (Atzori et al., 2011; Bermudez-Edo et al., 2016; Lu e Neng, 2010; Miao et al., 2010; Ning e Wang, 2011; Xu et al., 2014).
- A arquitetura do sistema IoT comprehende identificação por radiofrequência, redes de sensores sem fio, software de middleware, computação em nuvem e software de aplicativo IoT, conforme ilustrado no próximo slide (Lee & Lee, 2015).
- Esses aspectos das tecnologias de Big Data e IoT são descritos resumidamente a seguir.

Principais fontes de coleta de Big Data na IoT.



Identificação de rádio frequencia (RFID)

- A identificação por radiofrequência (RFID) permite a identificação automática e a captura de dados usando ondas de rádio, uma tag e um leitor.
- A tag pode armazenar mais dados do que os códigos de barras tradicionais.
- A tag contém dados na forma de um sistema global de identificação de itens baseado em RFID desenvolvido pelo Auto-ID Center (Khattab et al., 2017).
- As entradas do banco de dados para tags podem ter um tamanho efetivamente ilimitado.
- Portanto, o tamanho do banco de dados de uma tag e de seu objeto associado pode ser enorme (Juels, 2006).
- Por exemplo, em fábricas modernas, os processos usam recursos etiquetados com RFID. Esses recursos geram uma grande quantidade de dados logísticos enquanto se movem pelo processo de produção (Russom, 2011).
- A análise desta enorme quantidade de dados pode revelar informações e sugestões significativas na melhoria do planejamento logístico e layout de distribuição (Zhong et al., 2015).

Redes de sensores sem fio (WSN)

- Redes de sensores sem fio (WSN) consistem em dispositivos equipados com sensores autônomos e distribuídos espacialmente para monitorar condições físicas ou ambientais e podem cooperar com sistemas RFID para rastrear melhor o status de coisas como sua localização, temperatura e movimentos (Luigi Atzori et al.).
- Avanços tecnológicos em circuitos integrados de baixa potência e comunicações sem fio disponibilizaram dispositivos em miniatura eficientes, de baixo custo e de baixa potência para uso em aplicações WSN.
- Portanto, WSNs se tornaram um dos elementos mais importantes em IoT sendo uma tecnologia importante para apoiar a coleta de big data em ambientes internos onde podem coletar informações, por exemplo, sobre temperatura, umidade, condições de trabalho do equipamento, insumos de saúde e consumo de eletricidade.

Comunicações máquina a máquina (M2M)

- As comunicações máquina a máquina (M2M) representam um futuro onde bilhões de objetos do dia a dia e informações do ambiente circundante são conectados e gerenciados por meio de uma variedade de dispositivos, redes de comunicação e servidores baseados em nuvem (Wu et al., 2011).
- Mathew et al. (2011) descrevem uma arquitetura simples para a Web of Things (WoT), onde todos os objetos são conectados a um servidor baseado em conhecimento. O WoT é a versão inicial mais simples da IoT.
- O Bell Labs apresentou uma implementação de protótipo do WoT com quatro camadas: objetos físicos, um navegador WoT, lógicas de aplicativo e objetos virtuais (Christophe et al., 2011).
- Como os sensores M2M têm capacidade limitada de armazenamento e energia, suas redes requerem a transmissão de uma grande quantidade de dados em tempo real.
- Essa transmissão de dados precisa abordar as questões de eficiência, segurança e proteção (Suciu et al., 2016).
- Várias propostas foram apresentadas para resolver esses problemas na comunicação M2M. Por exemplo, canais de big data integrados à gestão do conhecimento foram propostos (Sumbal et al., 2017).

Computação em Nuvem

- A computação em nuvem é um modelo para acesso sob demanda a um pool compartilhado de recursos configuráveis (por exemplo, computadores, redes, servidores, armazenamento, aplicativos, serviços, software) que podem fornecer infraestrutura como serviço, software como serviço, plataforma como serviço ou armazenamento como serviço (Suciu et al., 2016).
- Consequentemente, os aplicativos IoT requerem armazenamento massivo de dados, uma alta velocidade para permitir a tomada de decisões em tempo real e redes de banda larga de alta velocidade para transmitir dados (Lee & Lee, 2015; Gubbi et al., 2013).

Big Data em áreas de aplicação de IoT - Sistemas de saúde

- IoT está oferecendo novas oportunidades para a melhoria dos sistemas de saúde ao conectar equipamentos médicos, objetos e pessoas.
- Desenvolvimentos tecnológicos associados a sensores sem fio estão tornando os serviços de saúde baseados em IoT acessíveis ainda mais longas distâncias físicas.
- Serviços de saúde baseados na Web ou eHealth às vezes são mais baratos e confortáveis do que a consultoria presencial convencional (Hossain & Muhammad, 2016; Sharma & Kaur, 2017).
- Computação em nuvem, Big Data e IoT e desenvolvimento de artefatos de TIC podem ser combinados na formação da próxima geração de sistemas de eHealth .
- O processamento de grandes quantidades de dados médicos heterogêneos, que são coletados de WSNs ou redes M2M, apóia um movimento de afastamento da pesquisa baseada em hipóteses em direção a pesquisas mais baseadas em dados.
- Outros programas, como motores de análise, extraem recursos e classificam os dados para auxiliar os profissionais de saúde a fornecer cuidados médicos adequados (Abawajy & Hassan, 2017).

Big Data em áreas de aplicação de IoT - Sistemas de saúde

- No campo da gestão clínica, os principais benefícios proporcionados por esses sistemas interativos incluem
 - (i) melhor tomada de decisão sobre o tratamento eficaz,
 - (ii) detecção precoce de erros no tratamento,
 - (iii) avaliação aprimorada do desempenho dos profissionais médicos ,
 - (iv) o desenvolvimento de novos modelos de segmentação e preditivos que incorporam dados de registro da unidade sobre o perfil dos pacientes,
 - (v) a automação do sistema de pagamento e controle de custos, e
 - (vi) a transmissão de informações para as pessoas certas no momento adequado.
- Ao reduzir drasticamente o tempo de armazenamento e processamento, as técnicas viáveis de Big Data também podem apoiar a atividade de pesquisa.

Cadeias de abastecimento de alimentos

- As cadeias de abastecimento de alimentos (FSC) existentes são processos muito complexos e amplamente dispersos que envolvem um grande número de partes interessadas.
- Essa complexidade tem criado problemas para a gestão da eficiência operacional, qualidade e segurança alimentar pública.
- As tecnologias IoT oferecem um potencial promissor para abordar a rastreabilidade, visibilidade e controlabilidade desses desafios no FSC (Gia et al., 2015; Xu et al., 2014), especialmente por meio do uso de tecnologias de código de barras e sistemas de rastreamento sem fio, como GPS e RFID em cada estágio do processo de produção, processamento, armazenamento, distribuição e consumo agrícola.
- Uma solução IoT típica para FSC compreende três partes:
 - dispositivos de campo, como nós WSN, leitores / tags RFID, terminais de interface de usuário, etc...
 - sistemas de backbone, como bancos de dados, servidores e vários tipos de terminais conectados por redes distribuídas de computadores, etc .; e
 - infraestrutura de comunicação como WLAN, celular, satélite, linha de energia, Ethernet, etc. (Xu et al., 2014).

Sistema de energia inteligente (Smart Grid)

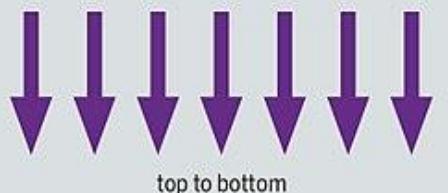
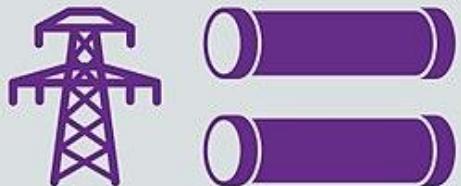
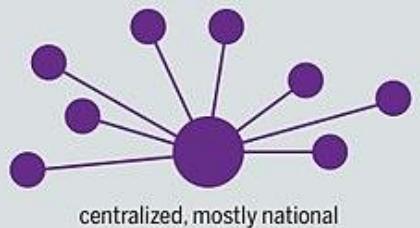
- Com os avanços na tecnologia IoT, sistemas inteligentes e análises de Big Data, as cidades estão evoluindo para se tornarem “mais inteligentes” (Stankovic, 2014).
- Por exemplo, os padrões das residências de uso de energia podem ser monitorados e analisados em diferentes períodos de tempo para gerenciar o custo da energia (Rathore et al., 2017).
- De acordo com pesquisas recentes, a tecnologia de smart grid é uma solução viável para ajudar a superar as limitações dos sistemas tradicionais de rede elétrica (Iyer & Agrawal, 2010; Parikh et al., 2010; Stojkoska & Trivodaliev, 2017).

Smart Grid

STAYING BIG OR GETTING SMALLER

Expected structural changes in the energy system made possible by the increased use of digital tools

yesterday



consumer

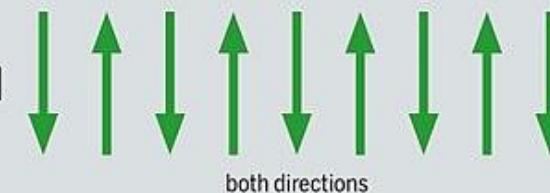
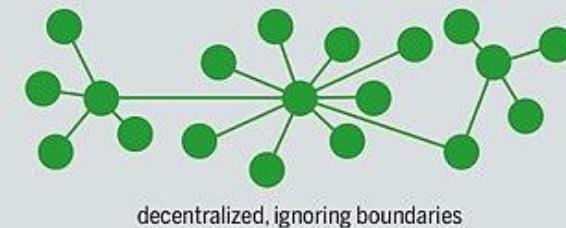
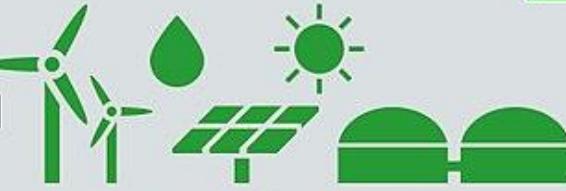
production

market

transmission

distribution

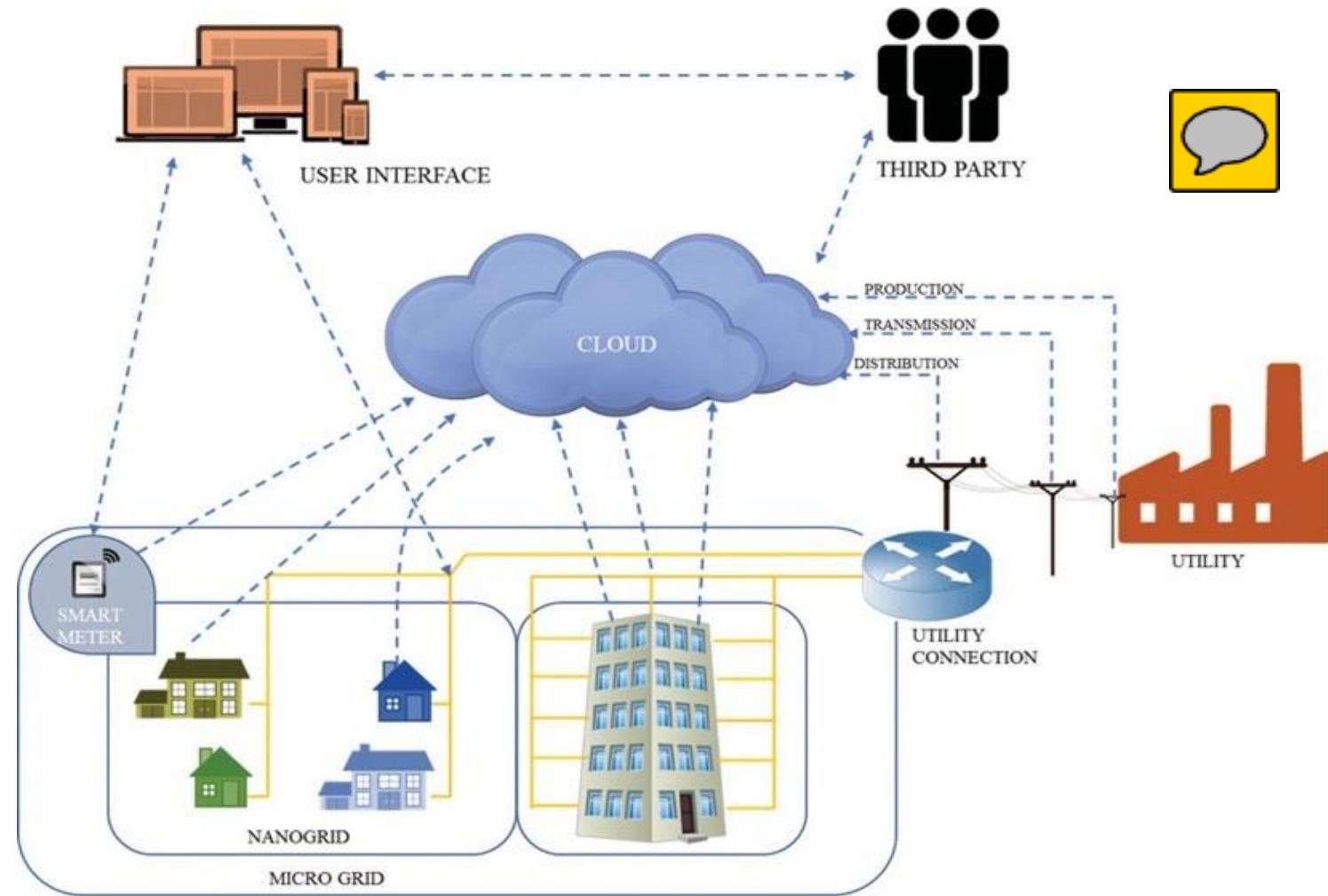
tomorrow



Casas inteligentes

- Stojkoska e Trivodaliev delinearam e propuseram uma estrutura generalizada para uma casa inteligente baseada em IoT (Stojkoska & Trivodaliev, 2017).
- Sua estrutura conecta residências, serviços públicos e provedores de aplicativos terceirizados por meio de uma rede em nuvem, com sensores conectados ao sistema de rede inteligente que coleta dados de eletrodomésticos inteligentes.
- Como a maioria dos serviços públicos aplica taxas de tempo de uso os fornecedores de aplicativos de terceiros podem reduzir os custos de serviços públicos combinando aparelhos como carregadores de bateria com geladeiras e fornos que podem ser controlados pela web (Buckl et al., 2009).
- Isso também se aplica a fontes de energia renováveis com medidores baseados na web que calculam quanta energia a casa exigirá da rede.
- A estrutura da casa inteligente é mostrada no próximo slide

Estrutura de IoT de vários níveis para casa inteligente (Stojkoska & Trivodaliev, 2017)





© 2014 iStockphoto.com