
Big Data Ecosystem
Seus dados nunca foram tão preciosos

Fábio Jardim

Formação:

- Bacharel em Ciência da Computação, UNIP
- Pós graduação em Análise de Big Data, FIA

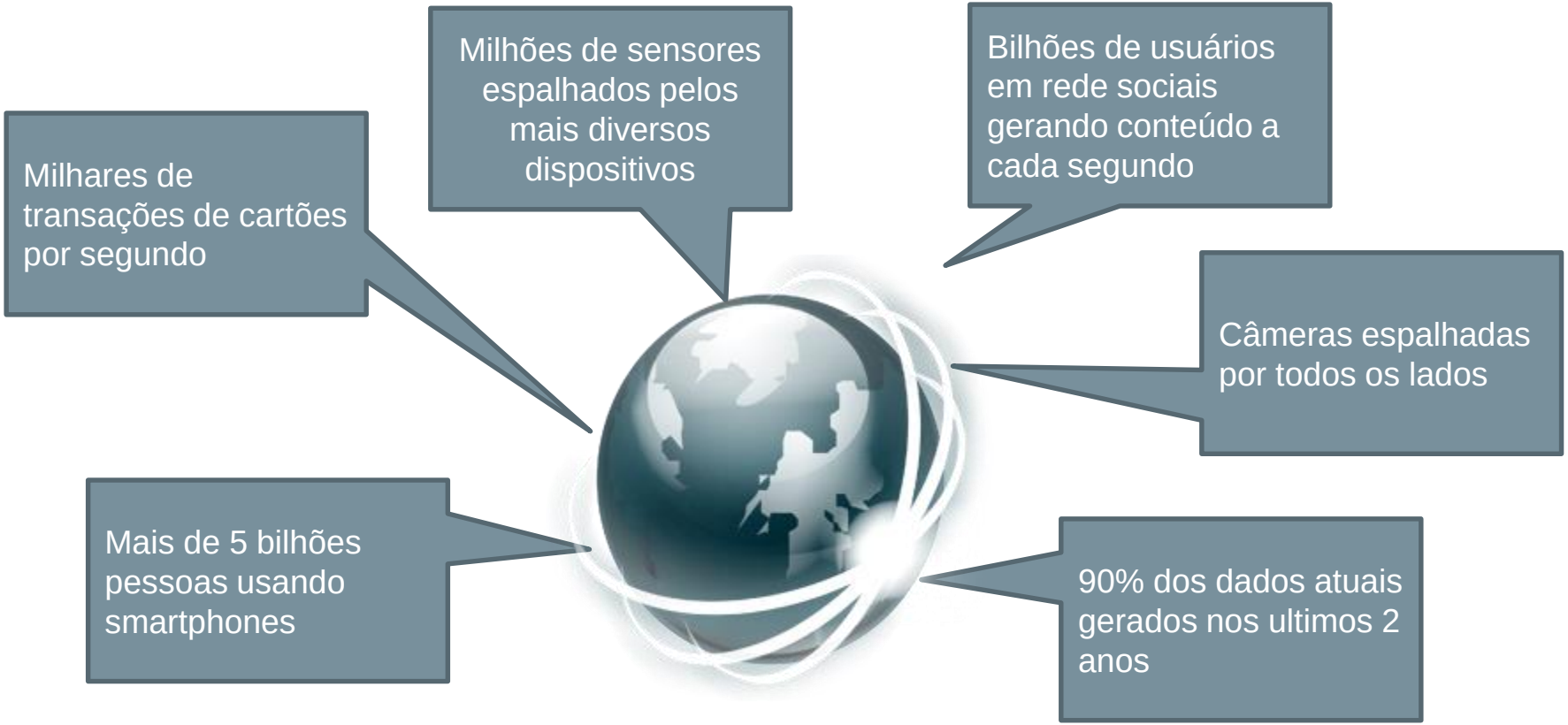
Experiência:

- Mais de 10 anos em plataforma de dados
- Vivência em grandes e-commerces como Magazine Luiza e Grupo Pão de Açúcar (Extra, Casas Bahia e Ponto Frio)
- Arquiteto Big Data na everis em projetos para grandes clientes
- Atualmente em projetos no setor bancário
- Professor Pós Graduação em Big Data na FIA

https://www.linkedin.com/in/fjardim/

fabiogjardim@hotmail.com

A Evolução dos Dados

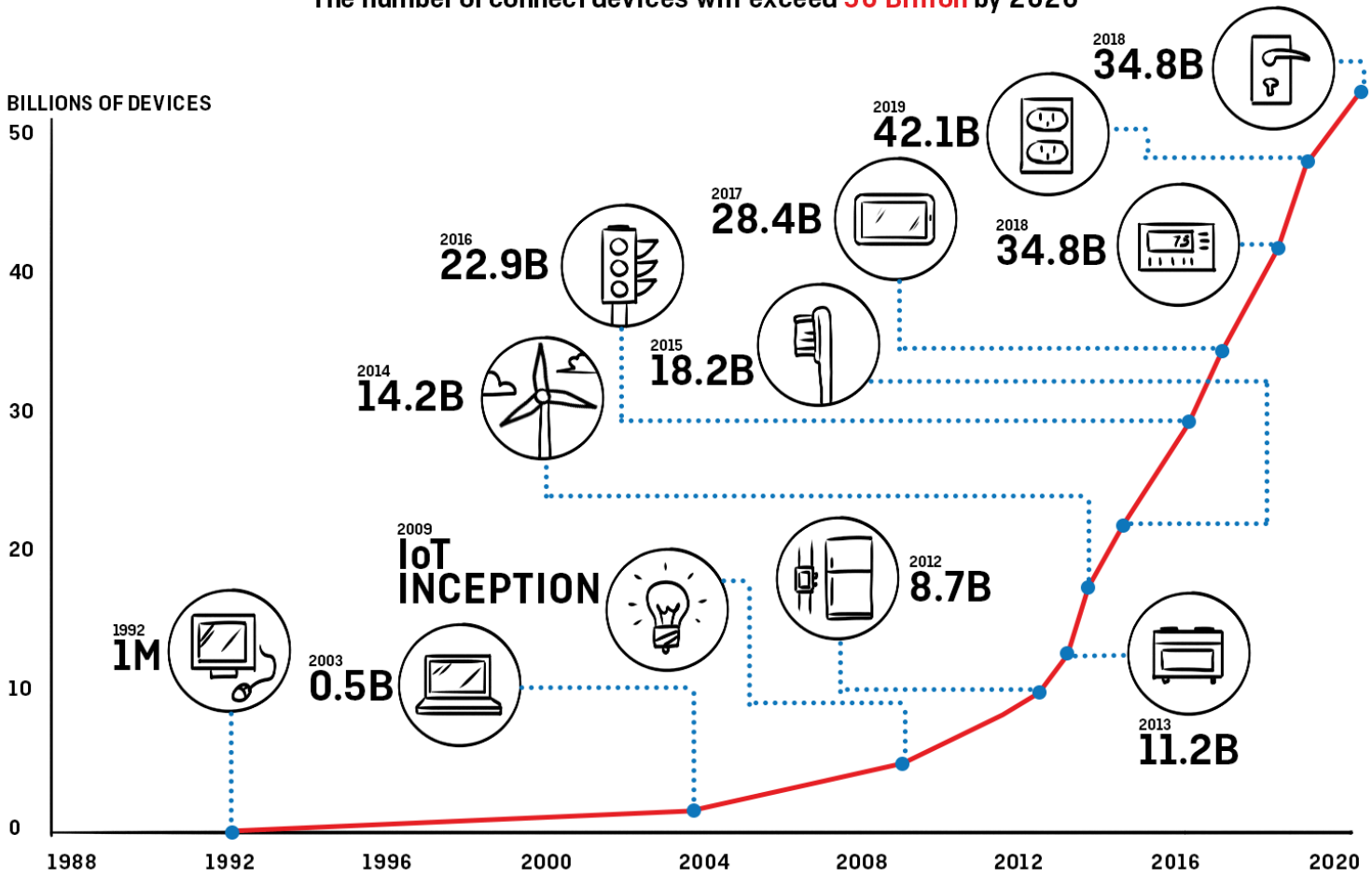


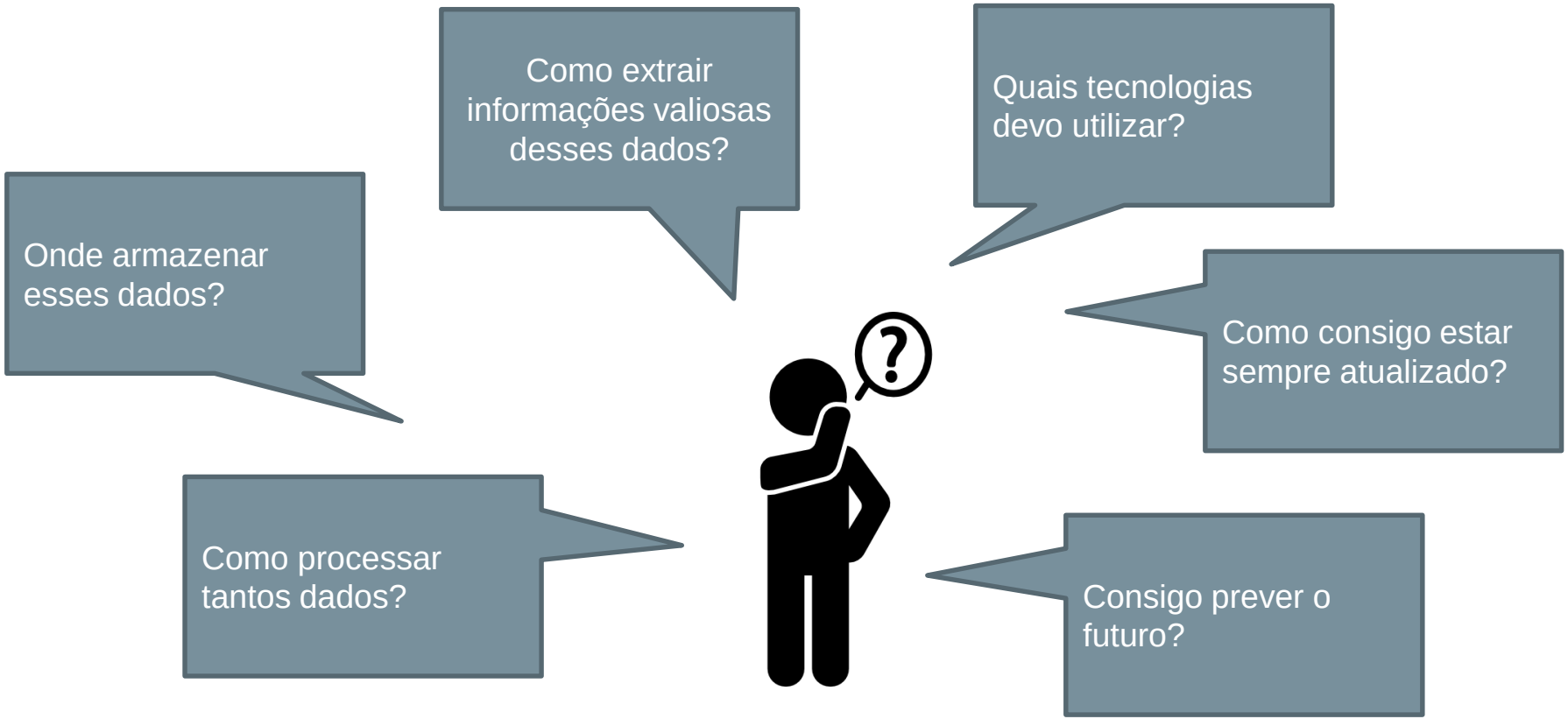
2017 This Is What Happens In An Internet Minute

Created By:
@LoriLewis
@OfficiallyChadd

Growth in the Internet Of Things

The number of connect devices will exceed 50 Billion by 2020




Como extrair informações valiosas desses dados?
Quais tecnologias devo utilizar?
Onde armazenar esses dados?
Como consigo estar sempre atualizado?
Como processar tantos dados?
Consigo prever o futuro?

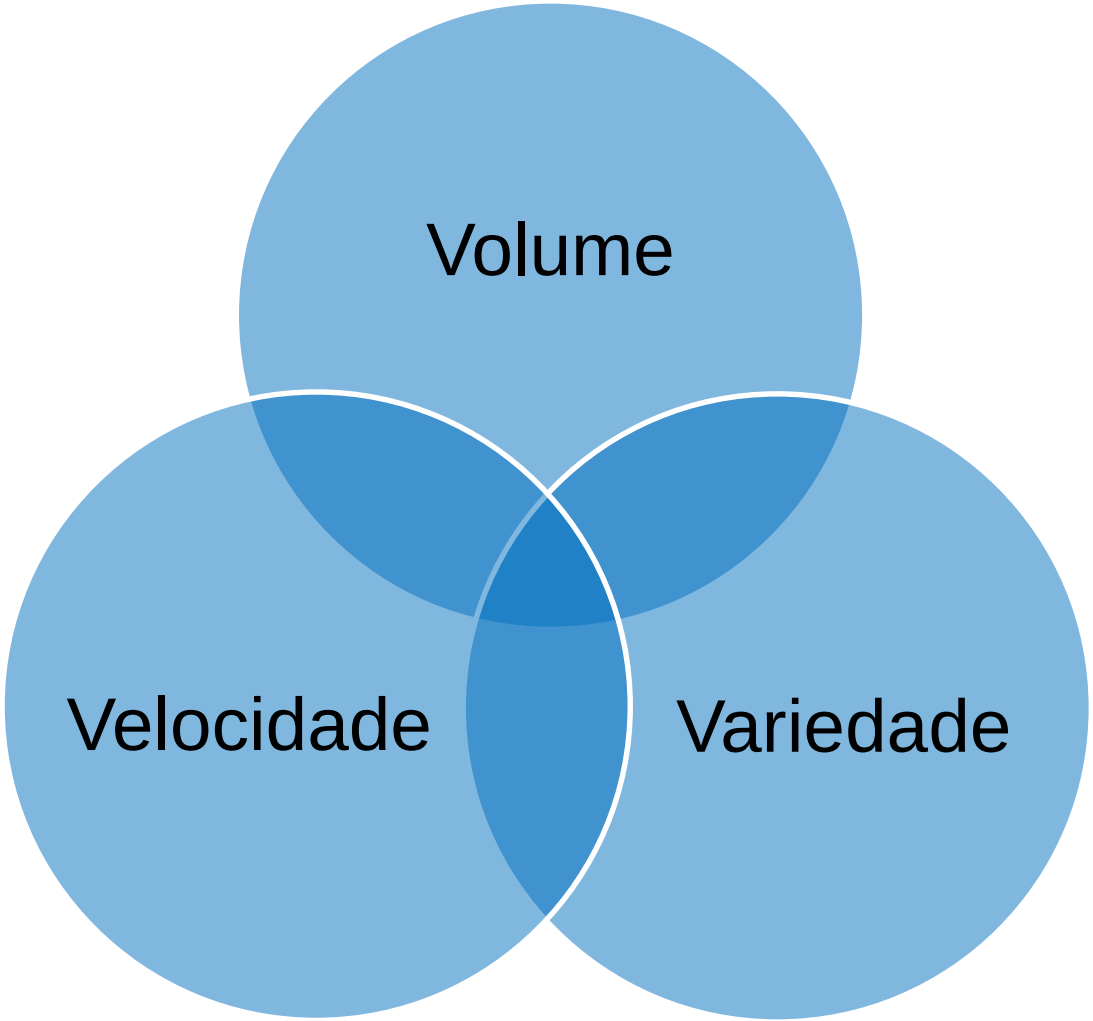
O que é Big Data

Grande conjunto de dados que excedem a capacidade de processamento de dados convencional.

Principais caratecristicas:

- Volume muito grande de dados
- Movem-se muito rápido
- Estruturados, semi-estruturados e não estruturados

Os Vs



Os Vs

As empresas armazenam
cada vez mais dados

Nunca fomos tão vigiados

Volume

A cada dia mais sensores
são introduzidos no nosso cotidiano

A Internet fez com que muito
mais dados fossem gerados

Os Vs

O tempo da informação gera lucro

O mundo esta conectado em Real Time

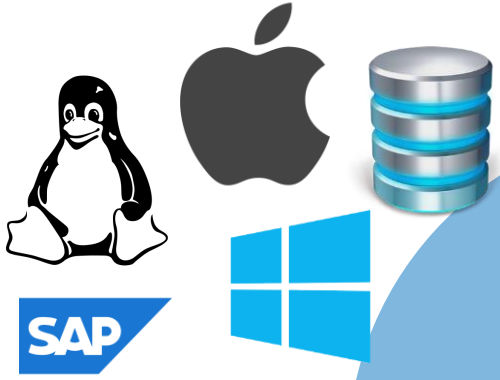
Velocidade

Ações são tomadas através de dados de sensores

A corrida dos dados

Os Vs

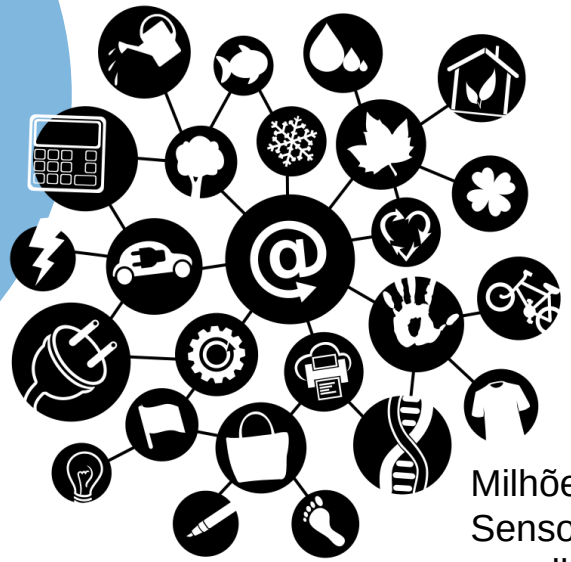
Dados coorporativos



Diversos dispositivos conectados

Variedade

Bilhões de usuários gerando conteudo em redes sociais



Milhões de Sensores espalhados

Onde armazenar tudo isso?

Nasce o conceito de Data Lake.

Conceito:

Vasto repositório com uma variedade de informações brutas que podem ser adquiridas, processadas, analisadas e entregues.

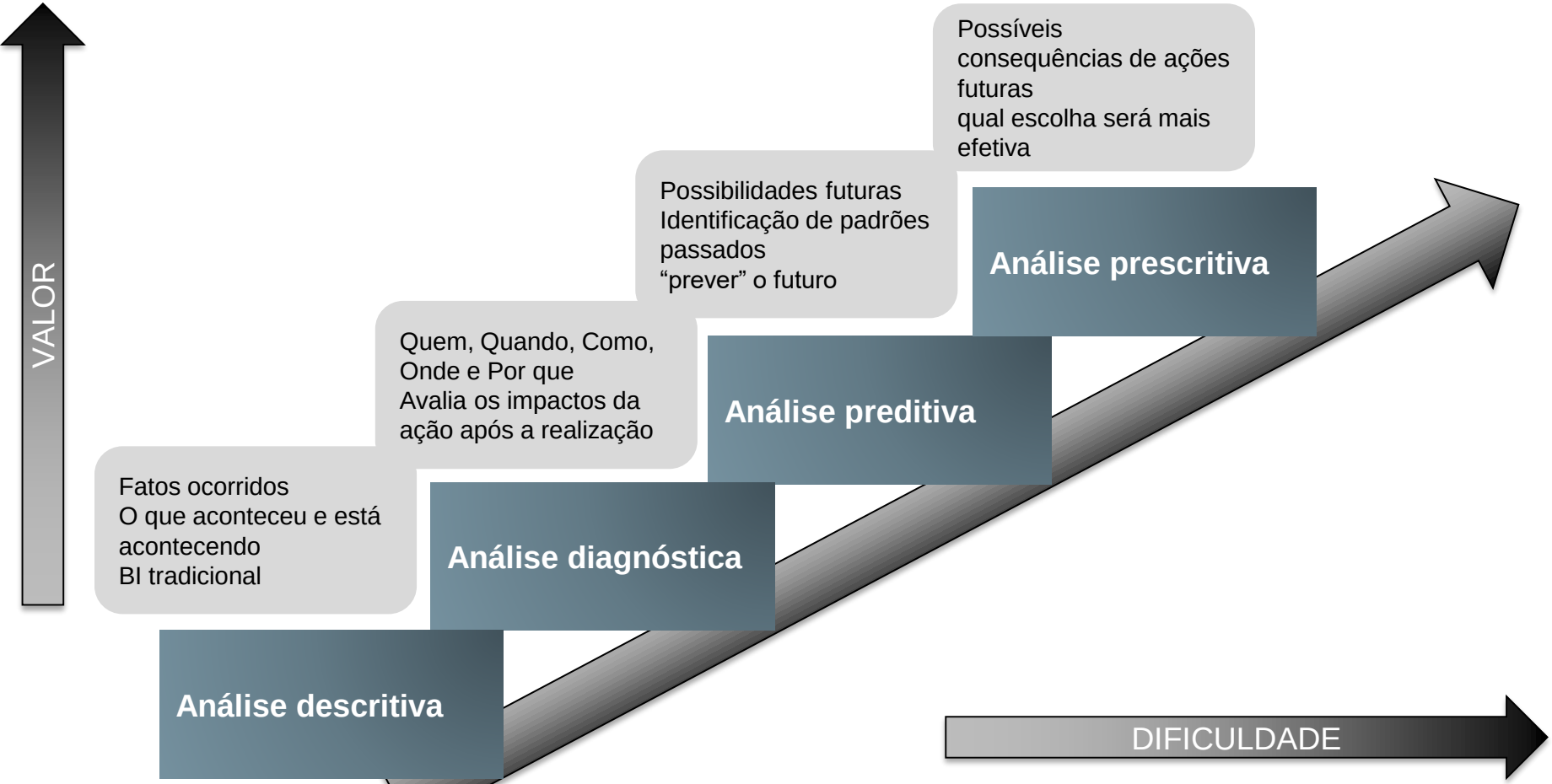
Propósito:

Derivar insights relevantes para a empresa a partir desta informação usando vários algoritmos de análise e aprendizagem de máquinas.

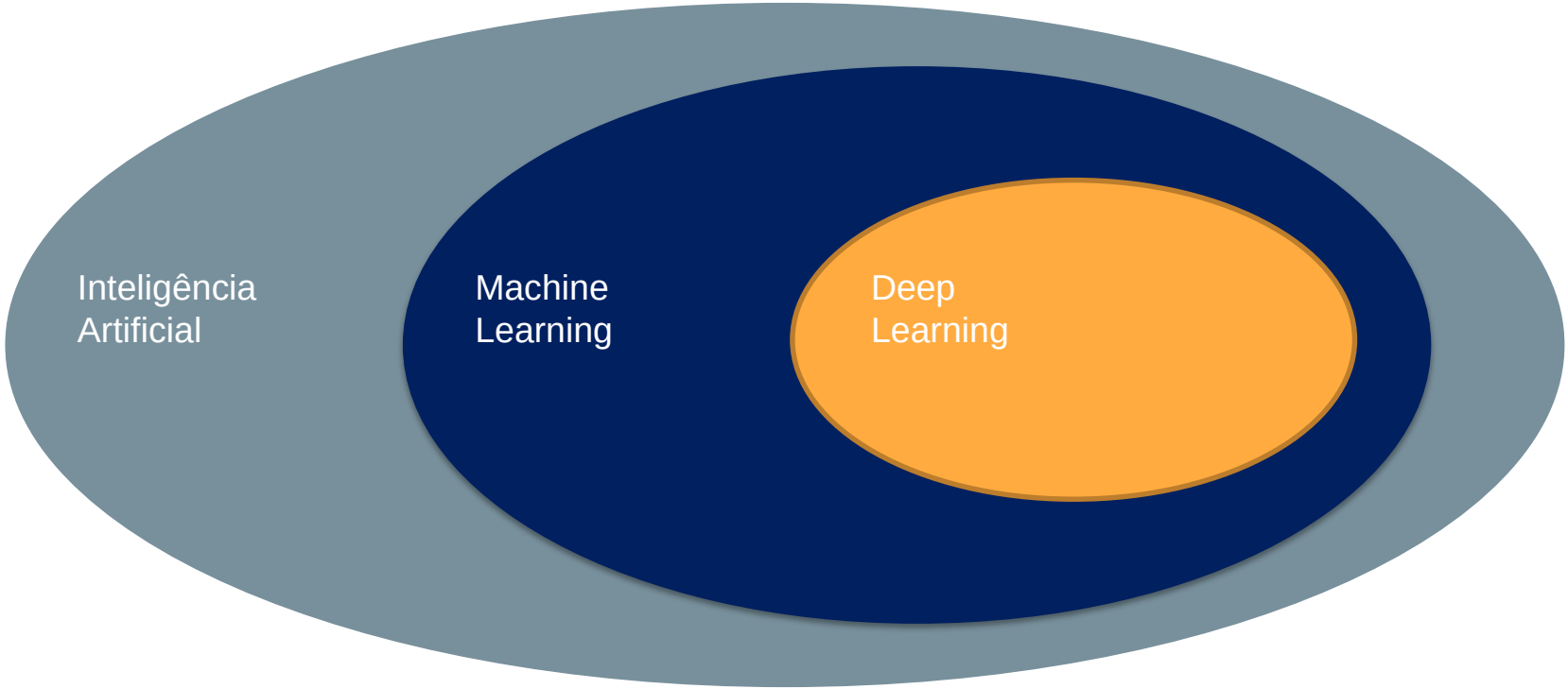
DW x Datalake

Data Warehouse	x	Data Lake
Estruturado e Processado	Dados	Estruturado, semi-estruturado e não estruturado
Dependente de esquema	Processamento	Livre de esquema
Alto custo para grandes volumes	Armazenamento	Desenvolvido para baixo custo
Configuração fixa, pouco agilidade	Agilidade	Configuração flexivel, alta agilidade
Consolidada	Segurança	Evoluindo
Área de negócios	Usuários	Data Scientists

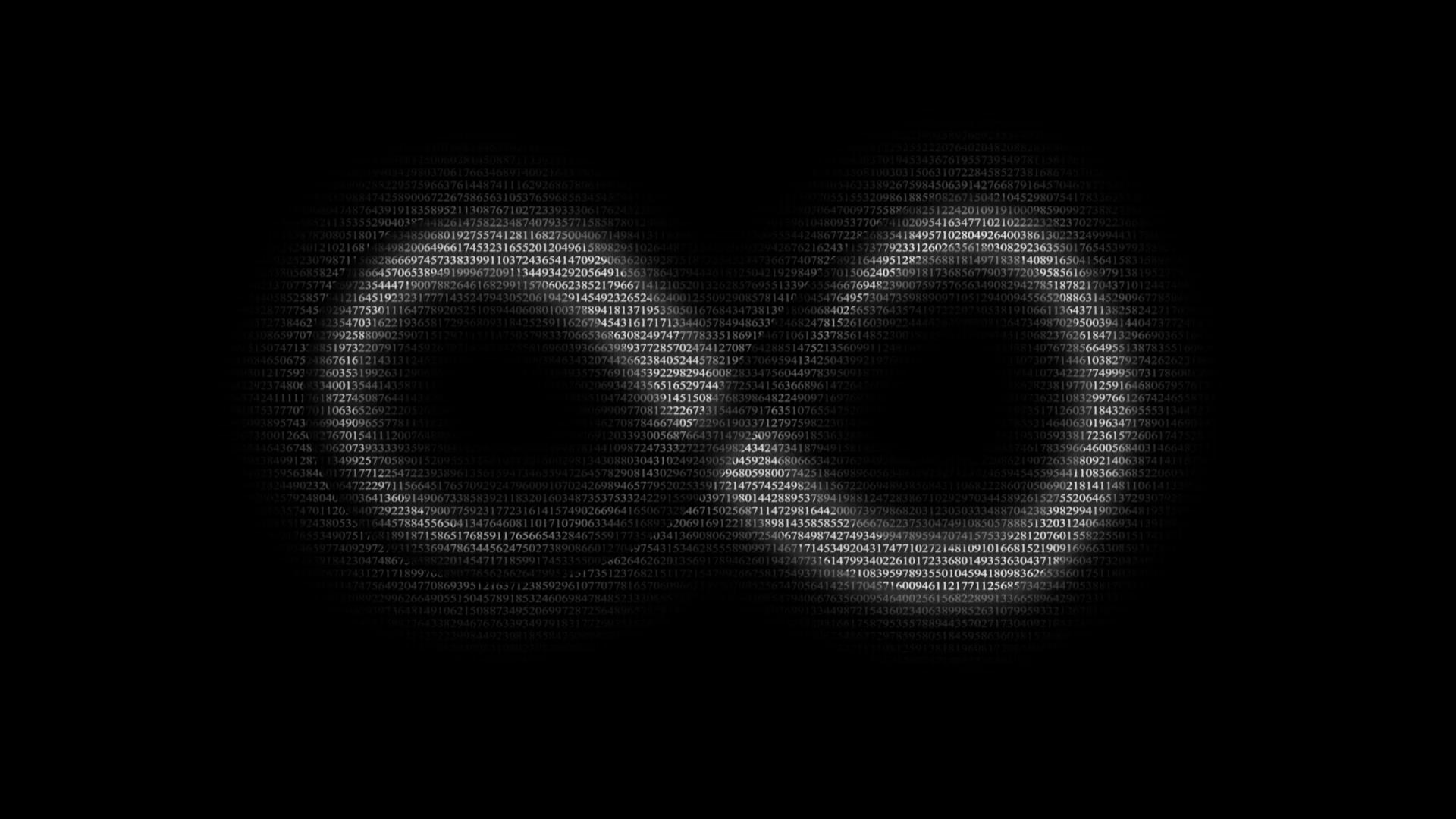
Como extrair valor dos dados?



Como extrair valor dos dados?




BIG
DATA
INFORMATION
STORAGE
LARGE
PETABYTES
SEARCH
ANALYSIS
MANAGEMENT
TECHNOLOGIES
GOVERNMENT
SOFTWARE
SETS
MILLION
VOLUME
PROCESS
DEVELOPMENT
TECHNOLOGY
SCIENCE
FUTURE
NEW
BILLION
PARALLEL
RESEARCH
INTERNET
LEARNING
WORLD
EVERY
BASED
TIMES
SAN
NAS
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
CAPACITY
HAND
ANNOUNCED
PAST
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
MAPREDUCE
CRITIQUES
STATE
CRITIQUE
PEOPLE
TRAFFIC
USE
CURRENTLY
APPROACH
MANY
OUT
ONE
FACTOR
MANAGE
ONLINE
DAY
WORLDWIDE
SYSTEMS
MAKING
INTEGRATION
MASSIVE
DEFINITION
USING
FRAMEWORK
BETWEEN
STATISTICS
BUSINESS
EFFECTIVE
VISUALIZATION
DATABASES
SIZE
ANALYTICS
SIMULATIONS
PROCESSING
EXABYTES
DETERMINE
WELL
SET
DEPARTMENT
QUERIES
HIGH
USERS
HIGHER
ALGORITHMS
COMPANIES
AMOUNT
RATE
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
TERMS
YEAR
PRIVATE
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LESS
HUGE
COMPARED
VARIOUS
SHARED
ARCHITECTURE
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
INTELLIGENCE
APPROACHES
PROCESSES
FLOW
INITIATIVE
ECONOMIC
TERABYTES
MARKET
VARIETY
RELATED
TIME
SENSOR
WOULD
PROJECT
DISTRIBUTED
PER
RELEVANT
ANNUAL
NOW
USED
PARADIGM
FUNDING
RECORDS
COST
INSIGHT
TOOLS
CENTER



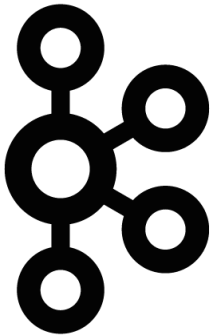
Case UPS Logística

- Mais de 4 bilhões de itens enviados por ano
- Frota de quase 100 mil veículos
- Otimização da frota
- Sensores nos caminhões e algoritmos avançados auxiliam com rotas, tempo ocioso dos motores e manutenção preventiva
- Economia de mais de 39 milhões de galões de combustível
- Evitou que seus motoristas dirigissem por 364 milhas desnecessárias

Case AMEX

- Mudança dos relatórios tradicionais e indicadores
- Modelos preditivos sofisticados
- Análise transações históricas
- 115 variáveis para prever o potencial de Churn.
- No mercado australiano, eles agora acreditam que podem identificar 24% das contas que fecharão dentro de quatro meses

Ecossistema Big Data



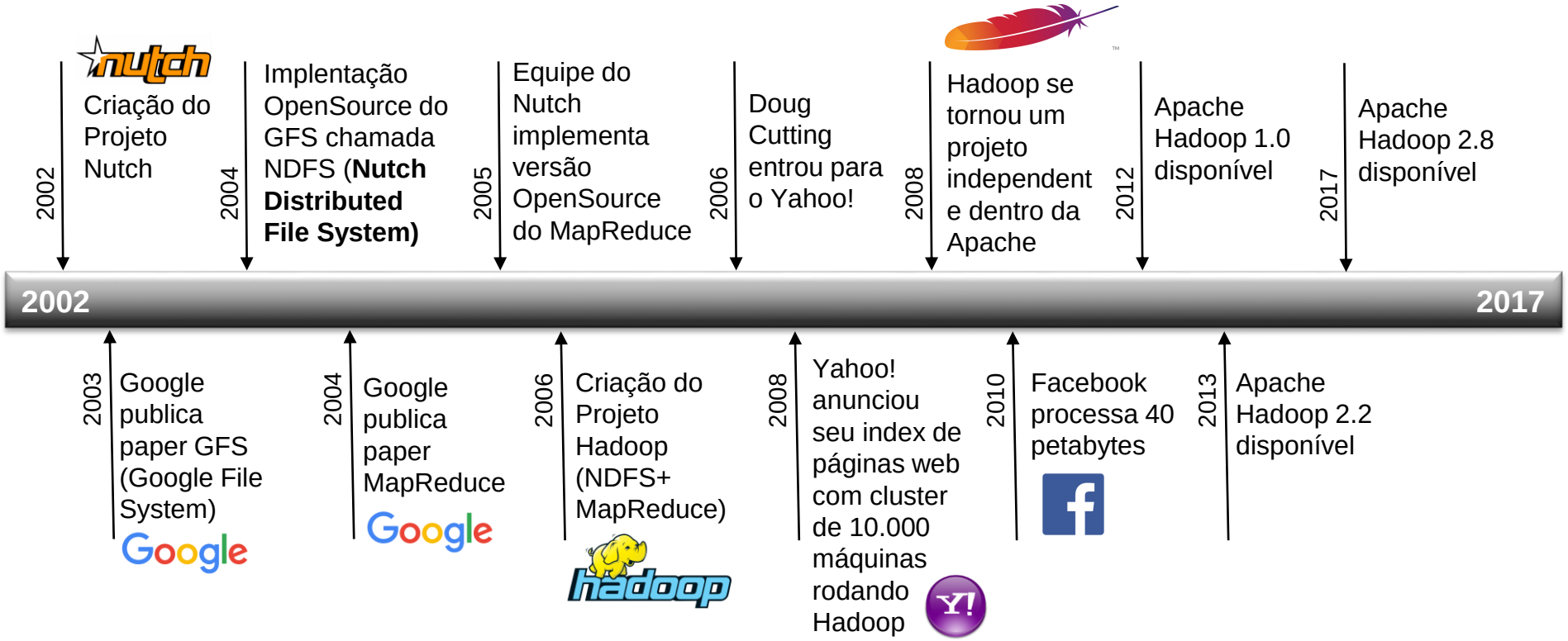
O que é Hadoop

- Plataforma que fornece infraestrutura resiliente, econômica e escalável
- Processamento em lote para grandes quantidade de dados
- Armazenamento e processamento distribuído
- Precursor do ecossistema Big Data
- Servidores commodities
- 4 módulos na versão 2.2
- HDFS, MapReduce, Hadoop Common e Yarn

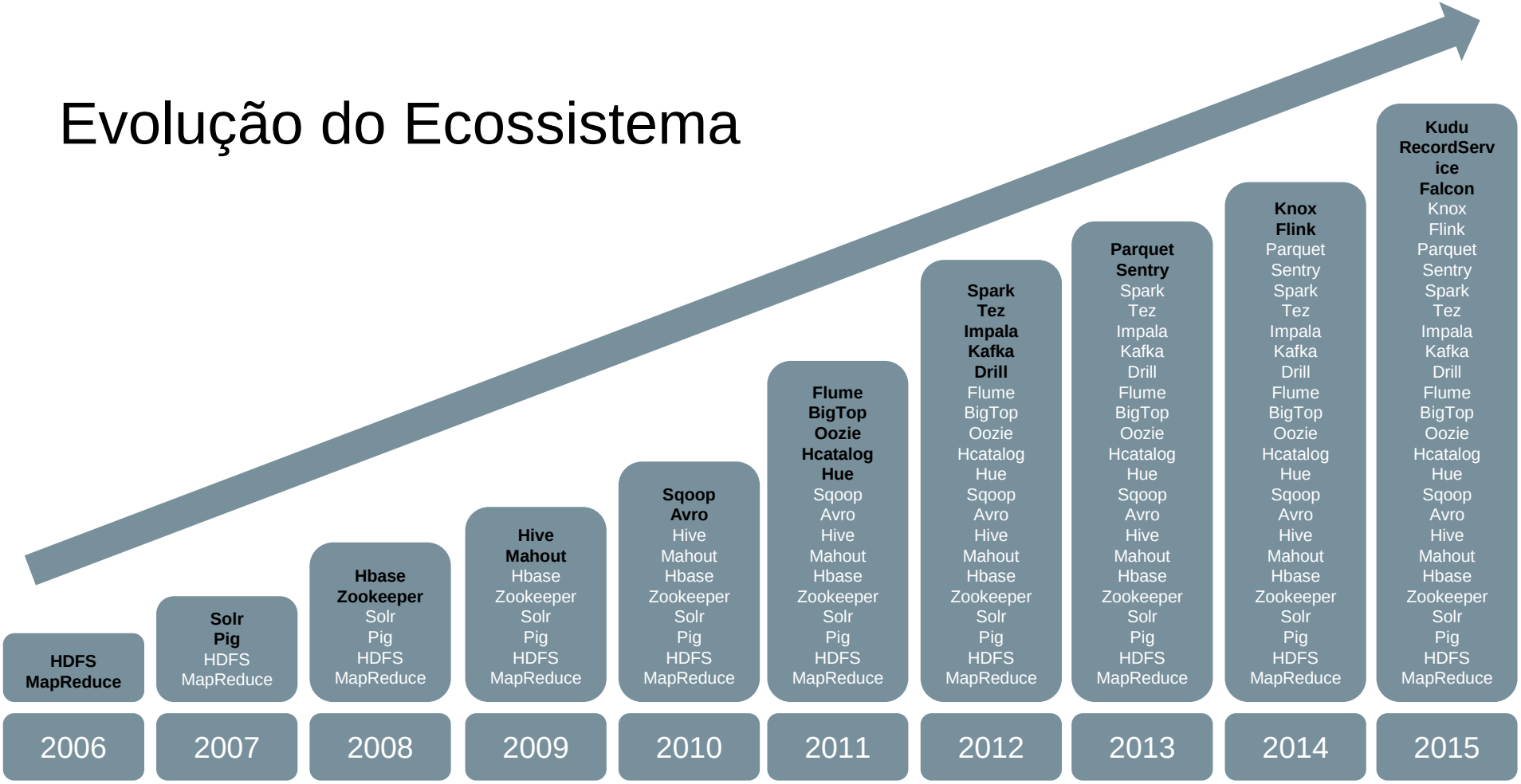
História

- Criado por Doug Cutting e Mike Cafarella
- Hadoop era o nome do elefante amarelo de pelúcia do filho de Doug
- Caminho para vários outros projetos, compondo o ecossistema Big Data

História


2002
Criação do Projeto Nutch
2004
Implentação OpenSource do GFS chamada NDFS (**Nutch Distributed File System)**
2005
Equipe do Nutch implementa versão OpenSource do MapReduce
2006
Doug Cutting entrou para o Yahoo!
2008
Hadoop se tornou um projeto independent e dentro da Apache
2012
Apache Hadoop 1.0 disponível
2017
Apache Hadoop 2.8 disponível
2002
2017
2003
Google publica paper GFS (Google File System)
2004
Google publica paper MapReduce
2006
Criação do Projeto Hadoop (NDFS+ MapReduce)
2008
Yahoo! anunciou seu index de páginas web com cluster de 10.000 máquinas rodando Hadoop
2010
Facebook processa 40 petabytes
2013
Apache Hadoop 2.2 disponível

Evolução do Ecossistema



Distribuições

cloudera®

MAPR

Comparativo

cloudera

Versão Free Limitada

Licença comercial

Alguns produtos proprietário

Interface amigável e completa de gerenciamento

Cloudera Impala

Gerenciador Cloudera Manager

Líder de mercado

Versão Free Ilimitada

Licença de suporte

100% Apache

no vendor lock-in

Gerenciado pelo Ambari, apenas básico

Por ser 100% Apache acompanha a evolução dos produtos mais rápido

Versão Free Limitada

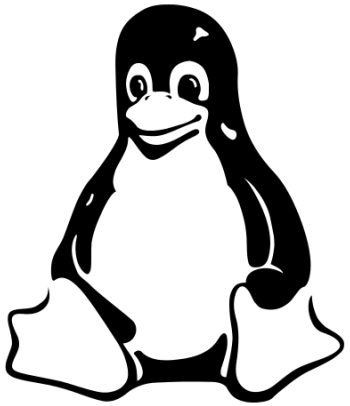
Licença comercial

Sistema de arquivos proprietário – MapRFS substitui o HDFS

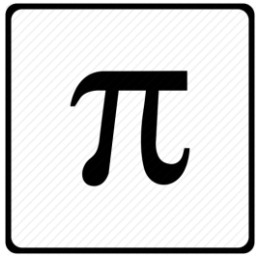
Considerada a distribuição mais rápida

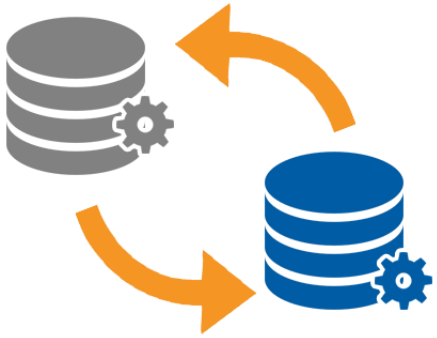
Gerenciador MapR Control System, não tem interface amigável

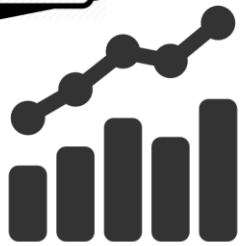
O que preciso saber?



NO SQL







Qual é o seu papel?

Data Engineer

Processamento Batch e Real Time

Consolidação de dados

Preparação dos dados para o Data Scientist

Estrutura de dados

Bando de dados relacional e NoSql

Conhecimentos: Hive, Python, Scala, HDFS, Spark, Hbase, Sqoop, Linuxe e shell, etc…

Big Data Architect

Definição de tenologia

Conhecimento abrangente entre as áreas

Conhecimento nos diversos frameworks, linguagens de programação e bando de dados

Conhecimentos : Hadoop, Spark, Storm, Kafka, Flume, Solr, Hbase, Pig, Hive, Zookeeper, Python, Java, Scala, Cassandra, Sqoop, Linux, Shell, cloud, network, etc…

Data Scientist

previsões

algoritmos de Machine Learning

responsável por aplicar técnicas de análise ao Big Data

extrair insights

Foco em modelos estatísticos

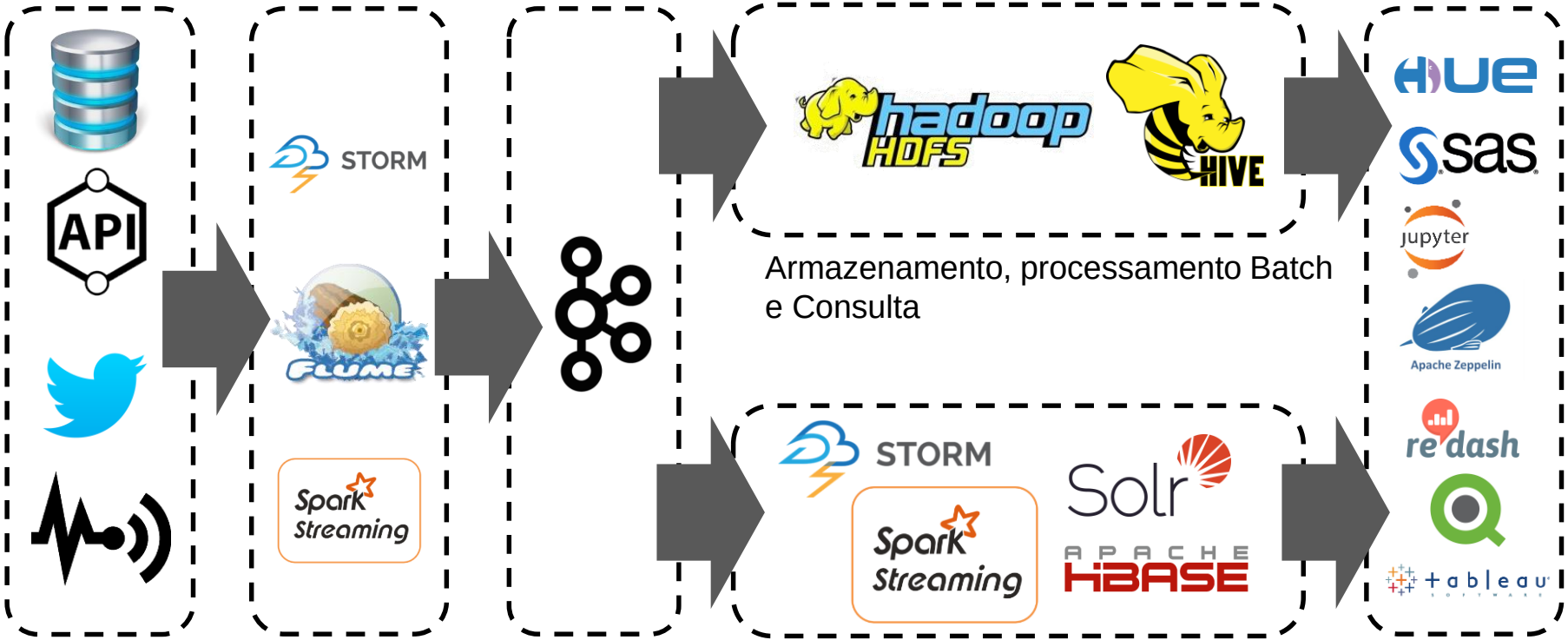
Conhecimentos: Conhecer o negócio, Estatistica,Scala, Python, R, SAS, SPSS, Machine Learning, Spark, Mahout, etc…

Big Data é legal

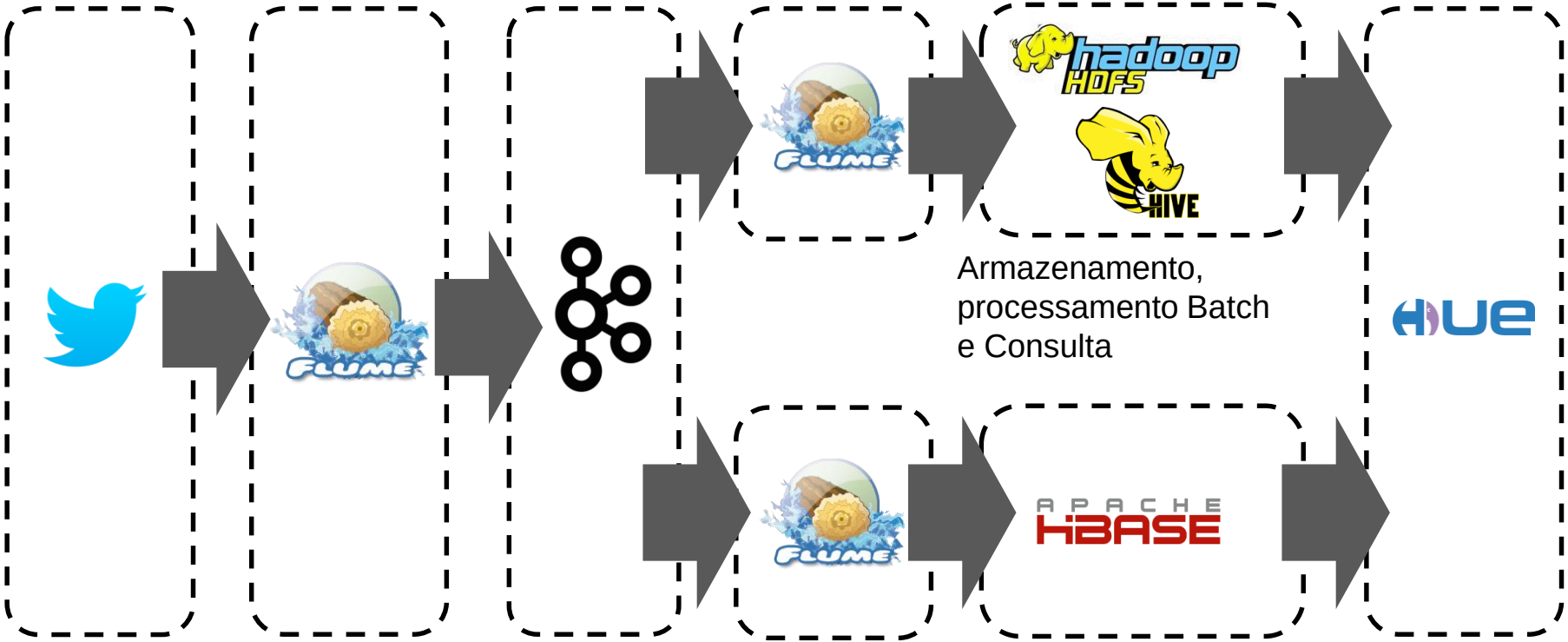
Mas….

Fast Data é muito mais legal

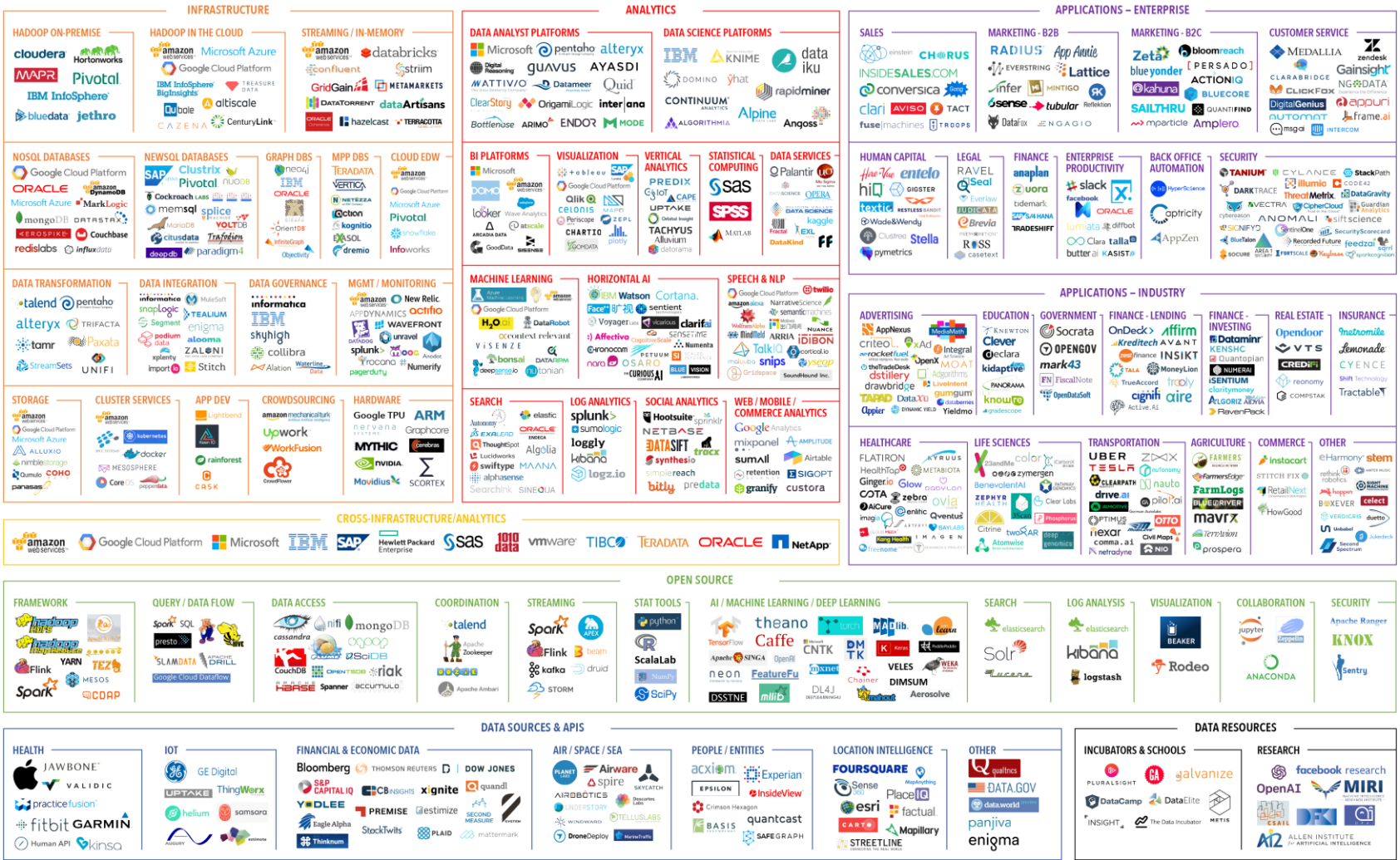
Algumas opções para aplicação Big Data Real Time



Exemplo aplicação Big Data Real Time



BIG DATA LANDSCAPE 2017



Meu banco de dados relacional vai morrer?

Quem pode usar Big Data?

Preciso utilizar Hadoop?

Por onde começar?

Posso ter mongoDb no meu Big Data?

Todos terão um Big Data?

Big Data é legal, mas Fast Data é muito mais

Tenho poucos dados, posso utilizar Big Data?

Posso subistituir meu BI tradicional por Big Data?