



# ARQUITETURA

BI Business Intelligence

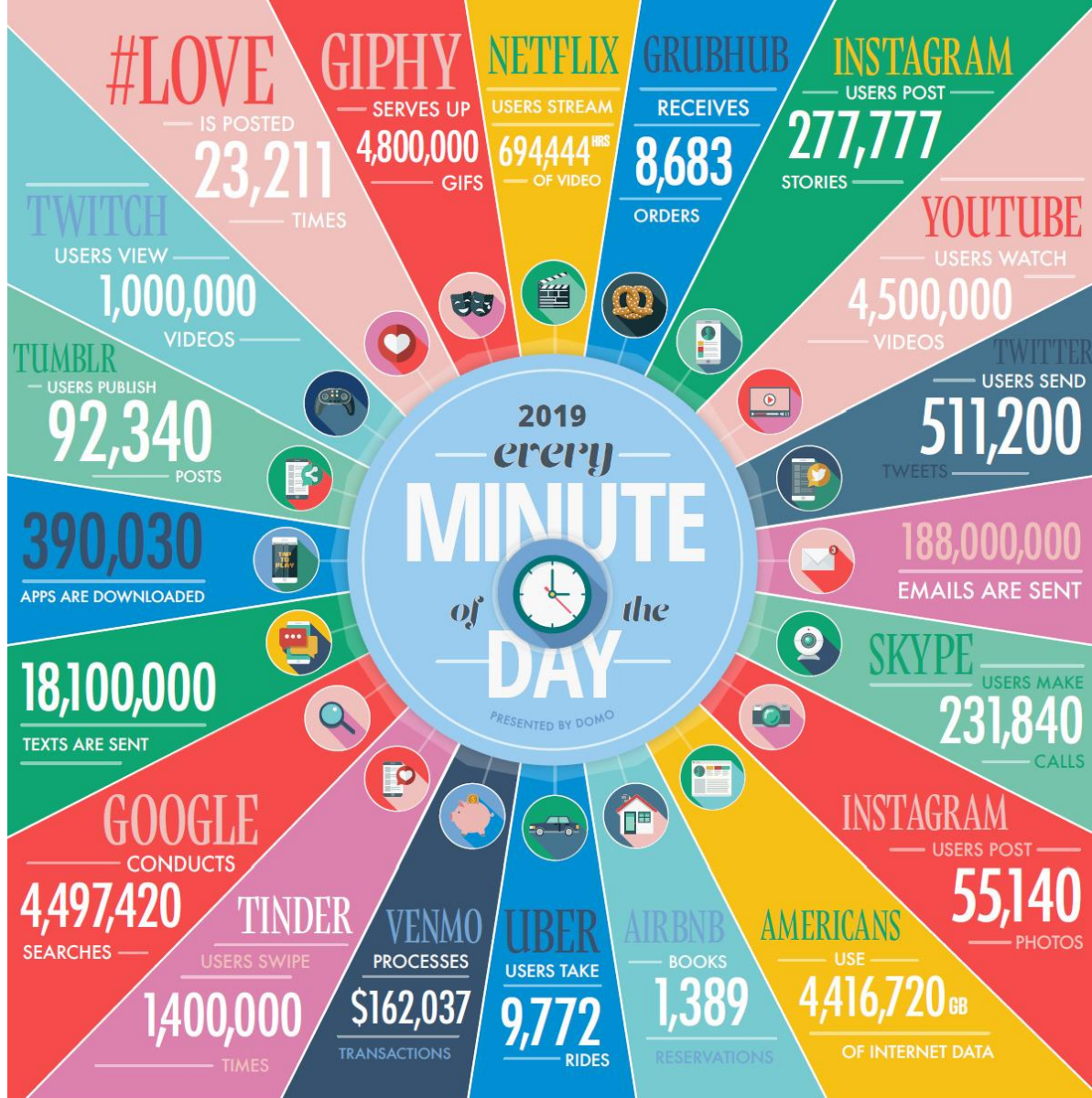
# CENÁRIO ATUAL CADA VEZ MAIS DADOS

Volume crescente de informações

Dados geram cada vez mais  
dados



As informações são vastas porém a sua  
percebibilidade tem a mesma velocidade  
de sua geração



# ONDE ESTÁ O PROBLEMA?



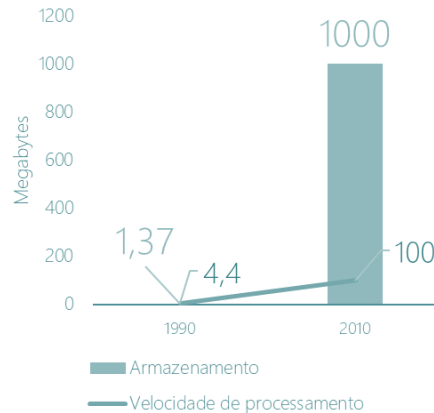
1990

Um típico disco poderia armazenar 1,370MB e tinha uma taxa de transferência de 4,4MB por segundo

*Deficiência em processamento*



Comparativo entre o armazenamento e o processamento



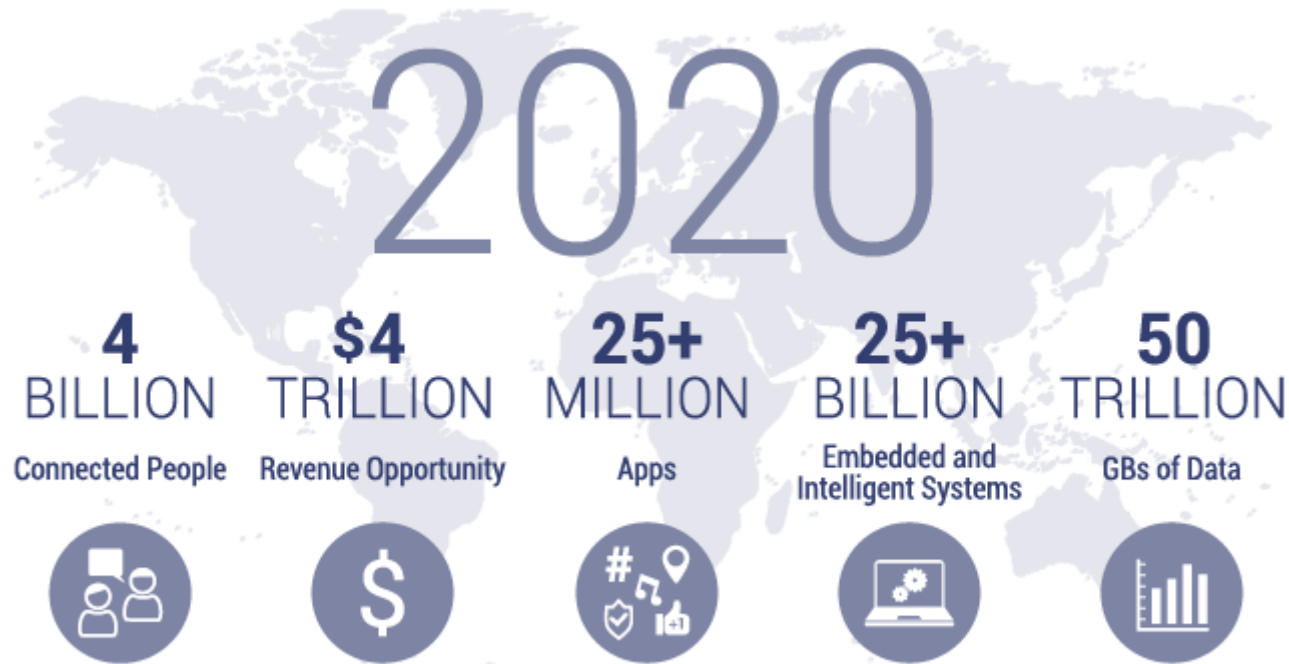
2010

Um típico disco poderia armazenar 1TB com uma taxa de transferência de 100MB por segundo

- ✓ Armazenamento cresceu aproximadamente 730 vezes;
- ✓ Processamento cresceu 23 vezes.

Fonte: especificações Seagate ST-41600n

# A IMPORTANCIA DO BIG DATA



Source: IDC

O principal objetivo do Big Data é poder tomar decisões **mais rápidas** e principalmente, com **maior acurácia**.

# BIG DATA – O QUE É?



## Volume

- Alto volume de informações geradas;
- Fontes internas e externas;
- 6 bilhões de usuários com celulares;
- Internet das coisas;
- 40 Zetabytes serão criados até 2020; 300 vezes mais que em 2005;
- 2.5 Quintilhões de bytes dia;



## Velocidade

- Percipibilidade do dado;
- Consumo imediato da informação;
- Sensores captam informações ex. Carro moderno mais de 100 sensores;
- 18.9 bilhões de redes conectadas;



## Variedade

- Texto, imagem, vídeo, áudio, logs, etc.
- Dados estruturados, desestruturados e semi estruturados;
- 4 bilhões de horas de vídeo no you tube;
- 7 petabytes de fotos/mês no facebook , 30 bilhões de conteúdos compartilhados por mês;
- 400 Milhões de tweets dia;



## Veracidade

- 1 em cada 3 líderes não confiam na informação disponível para tomada de decisão;
- Poor data custam U\$ 3.1 trilhão ao ano;
- 27% dos tomadores de decisão não sabem o quanto seus dados são incorretos;



# DIFERENÇAS ENTRE OS MODELOS



## Estruturado

---

- Linguagem única (SQL)
- Necessidade obrigatória de Schema
- Limitação de armazenamento
- Limitação de processamento
- Não volátil
- Arquitetura Cliente / Servidor
- Alta consistência
- Formato padronizado dos dados
- Fontes internas

## BigData

---

- Múltiplas linguagens
- Raw data desestruturado
- Armazenamento escalável
- Processamento escalável
- Dados voláteis
- Arquitetura de cluster
- Resiliente a falhas
- Diversos formatos
- Fontes internas e externas

# COMO FUNCIONA UM BIG DATA

O Big Data funciona através de um cluster

**Cluster é um conjunto de máquinas  
que se comportam como se fossem  
apenas uma**



Cada nó do cluster realiza simultaneamente as  
atividades abaixo:



## **ARMAZENAMENTO DISTRIBUÍDO**

- HDFS
- Data node
- Resiliente a falhas
- Map Reduce
- Spark
- Streaming

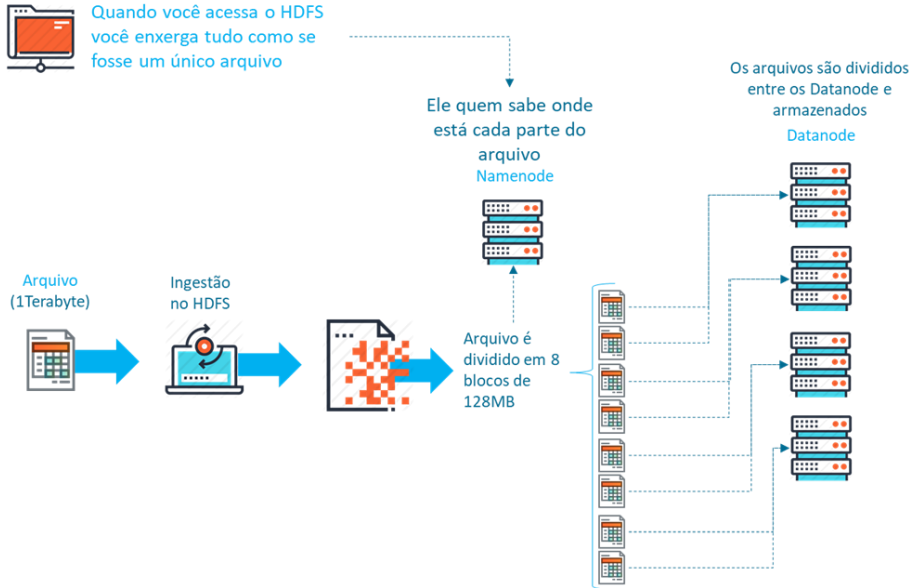


## **PROCESSAMENTO DISTRIBUÍDO**

- Yarn
- Job task
- Memória EMC
- RDD
- Spark
- Streaming



# IMPLEMENTANDO UM BIG DATA



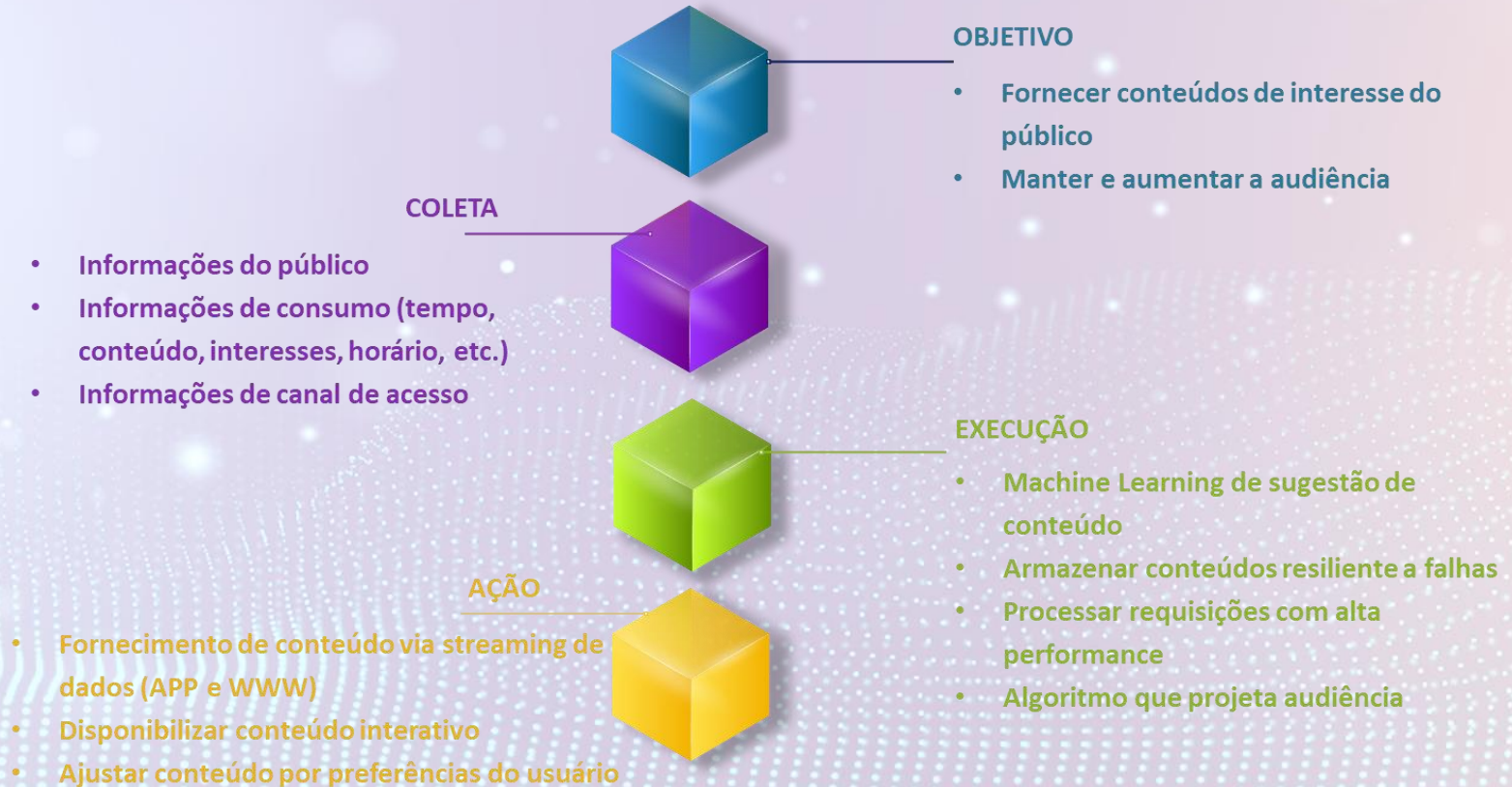
- Arquitetar o ecossistema
- Criar o Datalake
- Realizar a ingestão dos dados (Raw data)
- Integrar com outras estruturas de dados da organização
- *Criar camada manage*
- Extrair insights de bases de dados com elevado volume e complexidade
- Gerar Data products



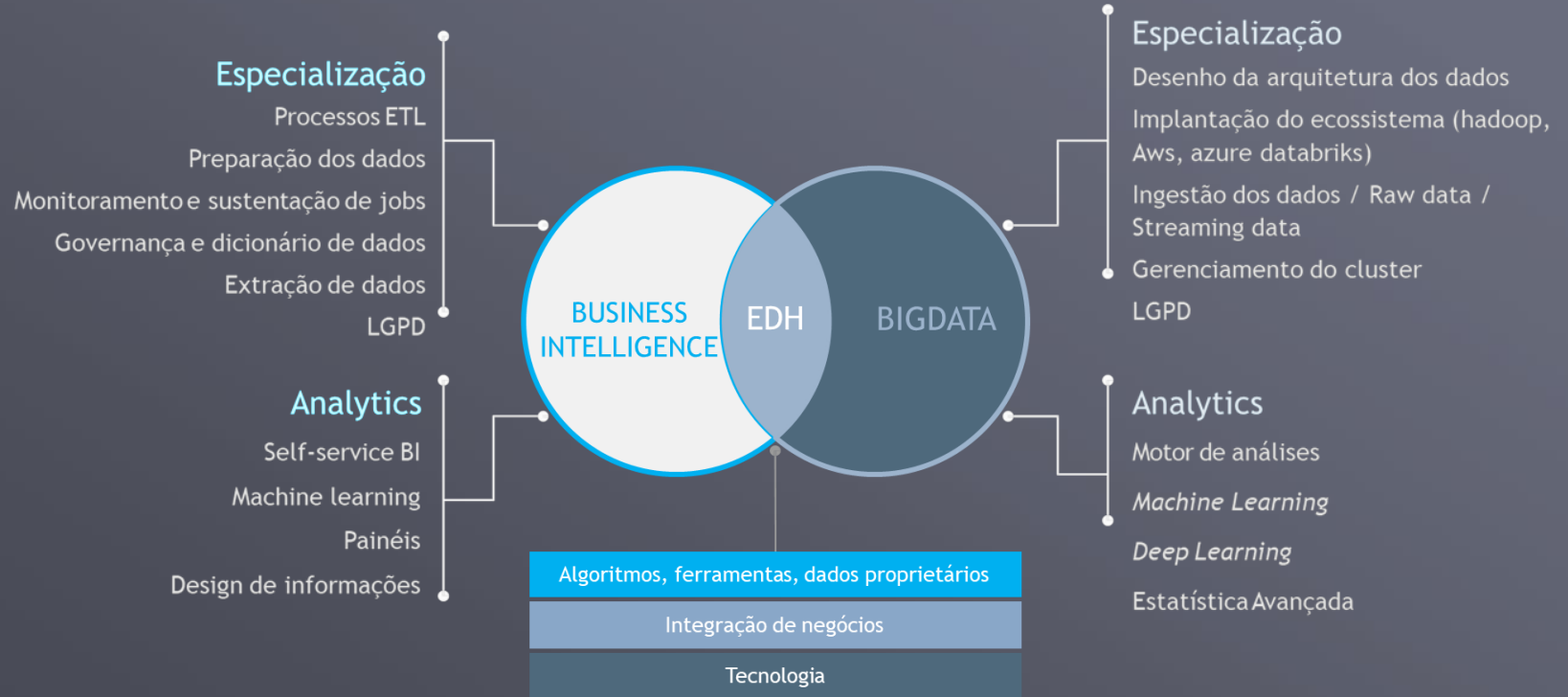
# APLICABILIDADES BIG DATA



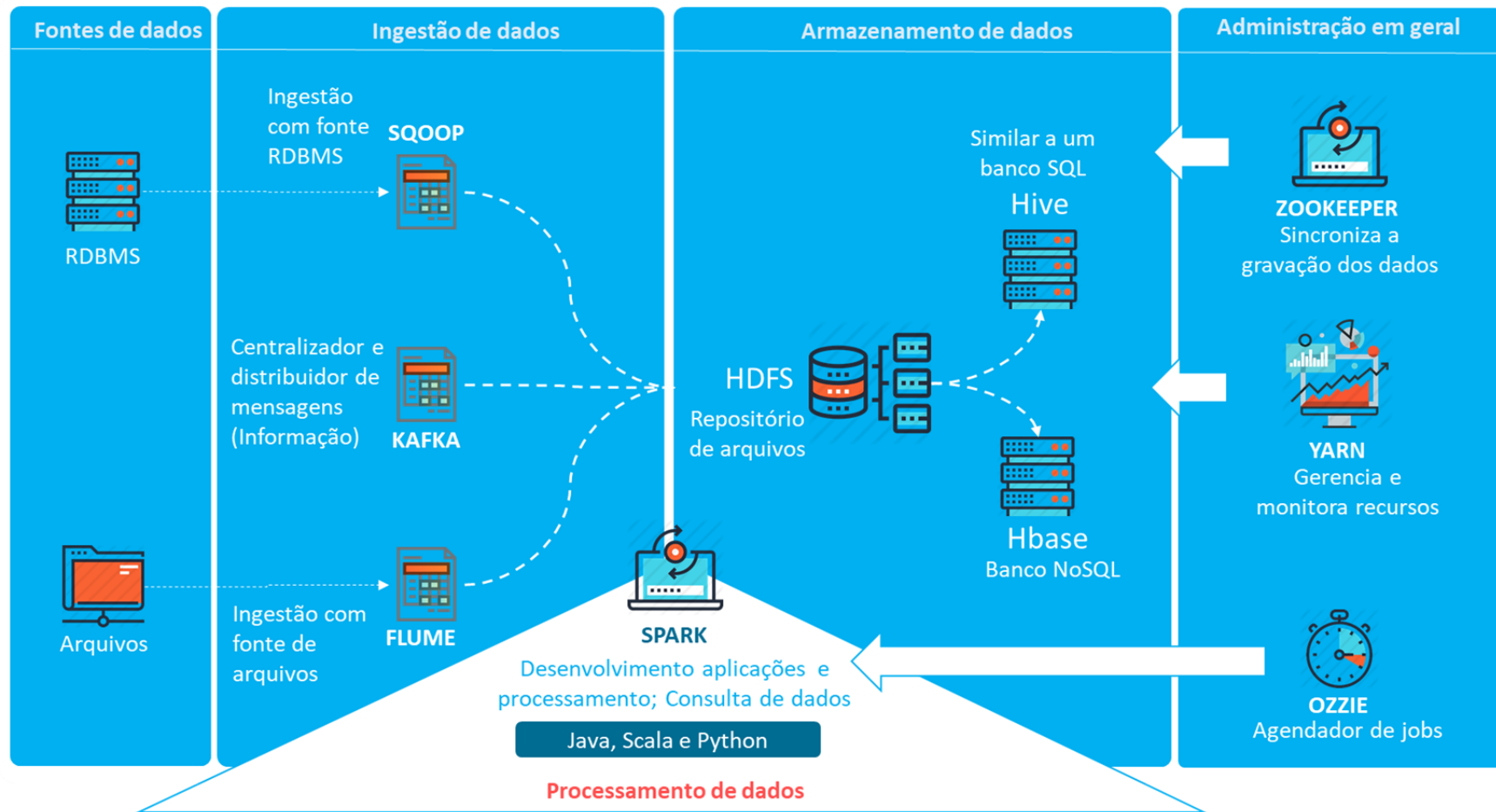
# APLICABILIDADES BIG DATA



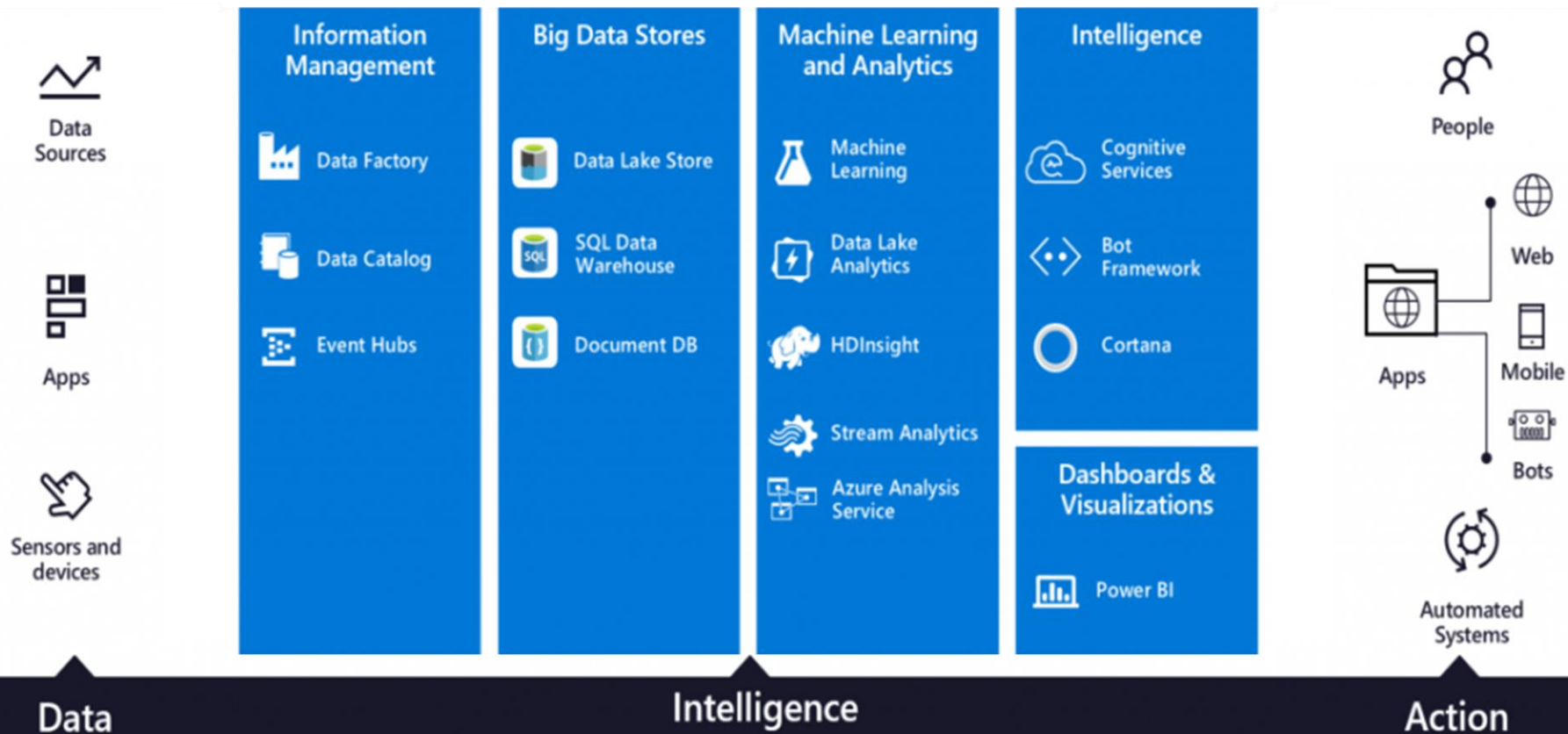
# BI & BIGDATA SOLUTION



# ECOSSISTEMA HADOOP



# ECOSSISTEMA MICROSOFT





# Portfólio de soluções



**ETL:** PowerCenter, Talend, NIFI, Pentaho, Data Service, SSIS, Data Stage, ODI...



**Bancos SQL:** SQL Server, Maria Db, Postgree, Oracle Database, Oracle Hexadata, MySQL, DB2, Teradata...



**Visualization:** Alteryx Designer, Tableau, OBIEE, QlickView, PowerBI, SBO, SSRS, Cognus, Grafana...



**Bancos NoSQL:** Mongo DB, Hbase, Cassandra, Impala...



**BigData:** HDFS, Kudu, MapReduce, YARN, Hive, Sqoop, Spark, Flume, Zookeeper, Ozie, Kafka, Tess, BDM, DataBriks, Azure, AWS...



**Machine Learning:** Mahout, Spark Mlib, Spark, Tensor Flow...



**Modelagem Preditiva:** Regressão / classificação Linear / não linear, Árvores de decisão, Inferência difusa, Bayesiano, Cadeias de Markov / séries temporais, Support vector machines



**Otimização:** Programação linear / inteira, Estocástico/ não linear, Desenho fatorial, Combinatória, Controle ótimo, Critérios múltiplos



**Simulação:** Simulação/ Monte Carlo, Martingale, Teoria de filas de espera, Teoria de jogos, Análise de dados topológicos, Análise linguista / Análise de imagem



**FIM**