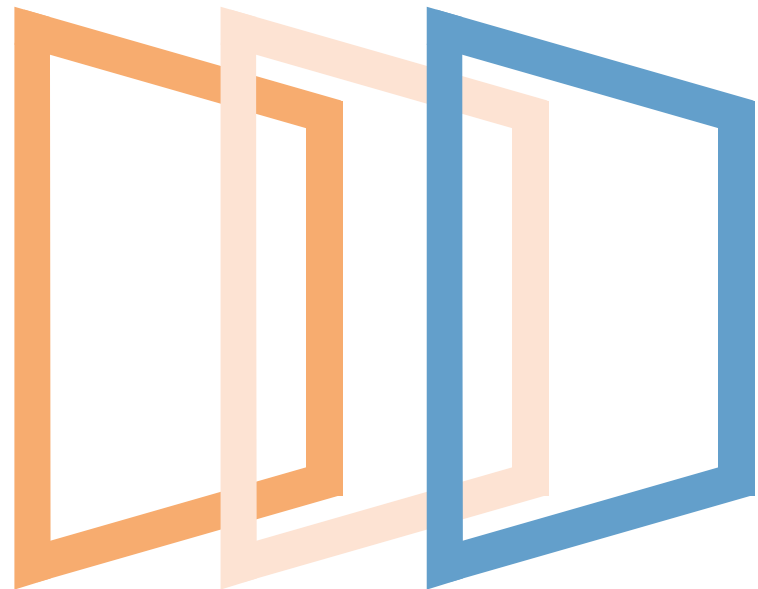


BI e Big Data

Aula 01 – Introdução ao BI e Big Data

minsoit



An Indra company

Sobre mim

Eu sou Caiuá França.

- Desenvolvedor Big Data Sr.
- Bacharel em Sistemas de Informação
- Pos Graduação em BI e Big Data

Minhas redes:

- <https://www.linkedin.com/in/caiuafranca/>
- <https://github.com/caiuafranca>

minsait



An Indra company

Índice

- Introdução ao Linux, Docker, Python, Git e Github

- O que é BI

- O que é Big Data

- Mercado Atual

- Arquitetura Hadoop

- HDFS

- Map Reduce

- Yarn

- Hadoop Common

minsoft

- Para onde a engenharia de dados esta indo

Linux, Git e Github, Python, Docker

Linux

<https://pt.wikipedia.org/wiki/Linux>

Git e Github

<https://pt.wikipedia.org/wiki/Git>

Python

<https://pt.wikipedia.org/wiki/Python>

Docker

[https://pt.wikipedia.org/wiki/Docker_\(software\)](https://pt.wikipedia.org/wiki/Docker_(software))

O que é BI?

Vamos partir do básico. A **inteligência de negócios** ou **business intelligence** é mundialmente conhecida pela sigla **BI**.

Inicialmente, focando no termo **intelligence** ou **inteligência**, fica mais fácil de você começar a entender que o **BI** não é “alguma coisa”, mas sim um “conjunto de coisas”.

E este é o nosso ponto de partida: que coisas?

O foco do BI é responder as perguntas que o negócio espera

Fundamentos do BI?

Fundamentalmente, elas são:

pessoas

dados

processos

tecnologias

ferramentas

sistemas

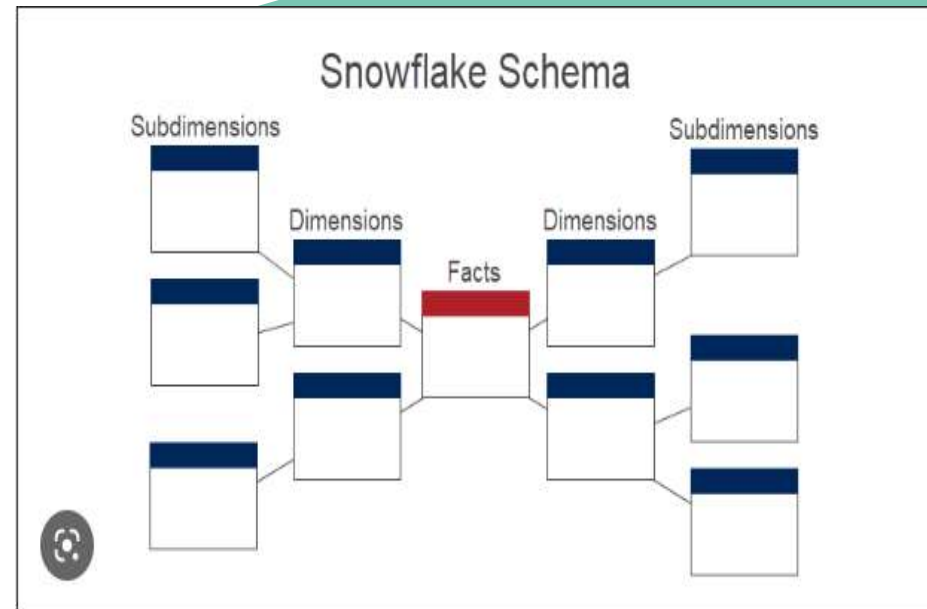
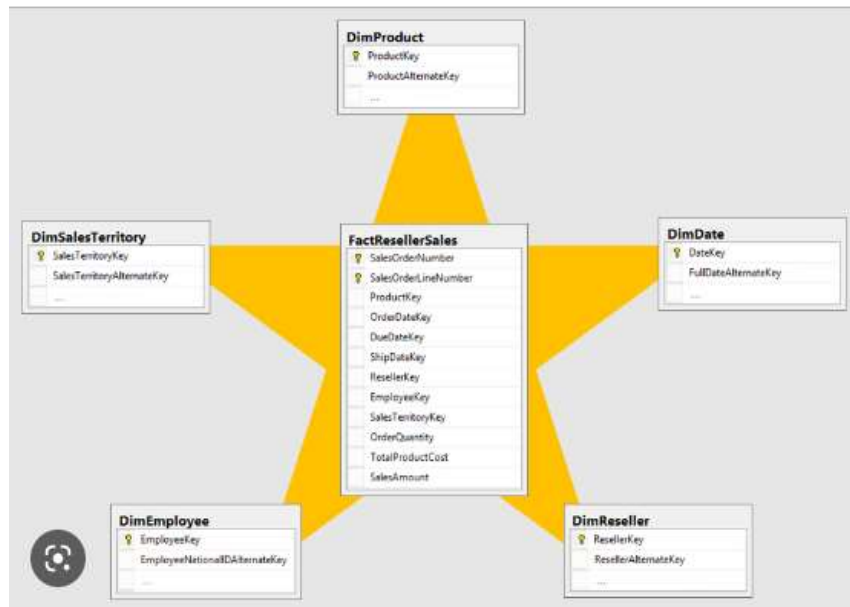
cultura organizacional

Modelagem de Dados

Modelagem dimensional é uma técnica de projeto lógico normalmente usada para data warehouses que contrasta com a modelagem entidade-relacionamento.

Segundo o prof. Kimball, a modelagem dimensional é a única técnica viável para bancos de dados que devem responder consultas em um data warehouse. Ainda segundo ele, a modelagem entidade-relacionamento é muito útil para registro de transações e para fase de administração da construção de um data warehouse, mas deve ser evitada na entrega do sistema para o usuário final

Tipos de Modelos



Elementos do Modelo

Fatos:

A tabela de fatos, no "centro" da estrela, fica rodeada por tabelas auxiliares, chamadas de tabelas dimensão. A tabela de fatos conecta-se as demais por múltiplas junções e as tabelas de dimensões se conectam com apenas uma junção a tabela de fatos.

Dimensões:

A dimensão é tudo que qualifica o fato que estamos analisando
Ex. tempo, localidade e etc..

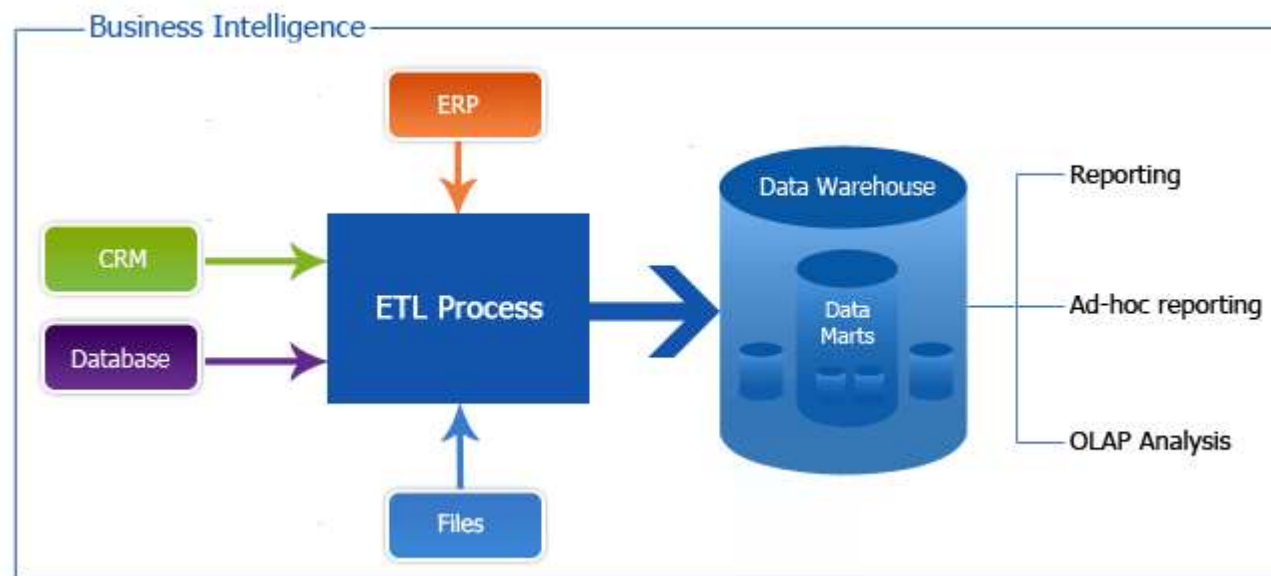
Granularidade:

É até que nível a minha análise deverá ir
Ex. minha análise deverá ir até o dia da compra ou a hora que a compra foi feita?

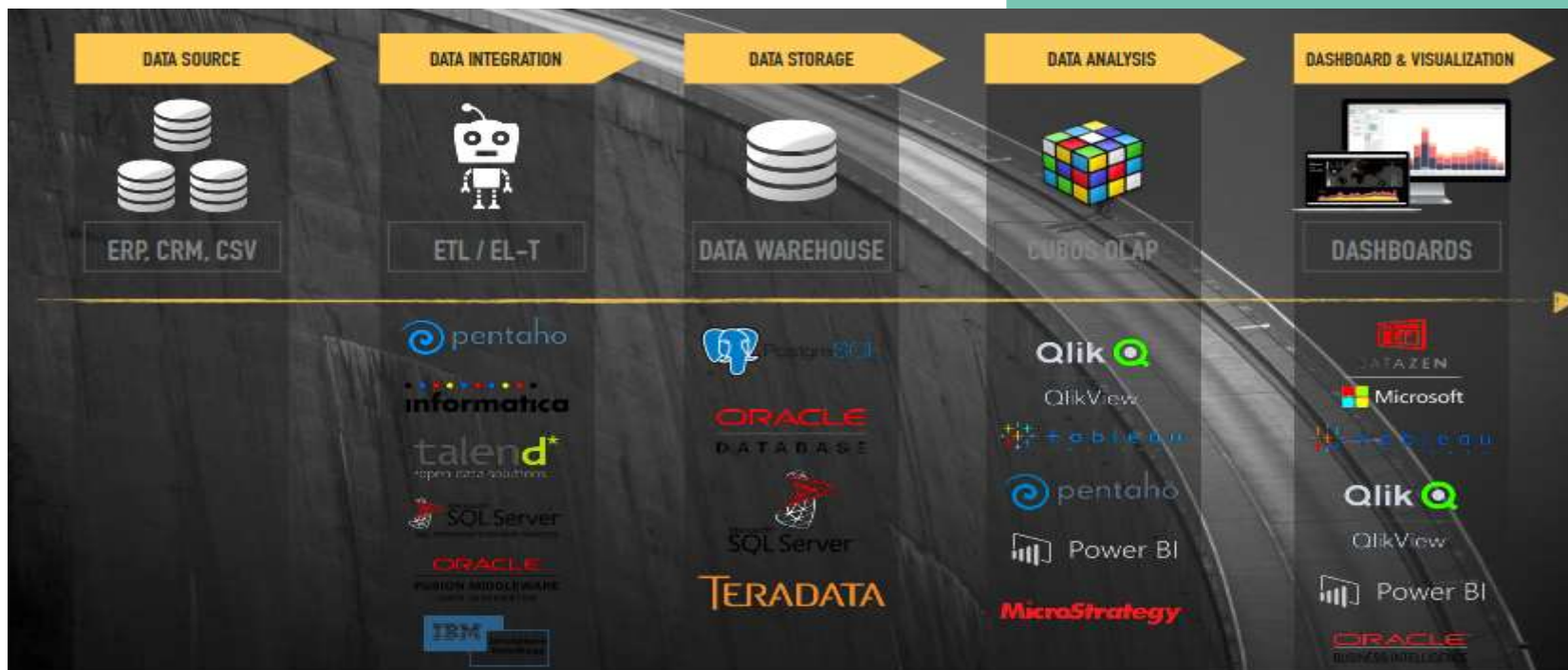
DW ou Data warehouse

Conceito:

https://pt.wikipedia.org/wiki/Armaz%C3%A9m_de_dados



Etapas do BI?



O que é Big Data?

São dados com maior volume, velocidade e variedade, que softwares tradicionais de processamento não conseguem gerenciar. (oracle.com)

São informações de alto volume, velocidade e variedade que exigem formas inovadoras e econômicas de processamento e permitem uma visão aprimorada, para tomada de decisões e automação de processos. (gartner.com)

Os 5 v's do Big Data

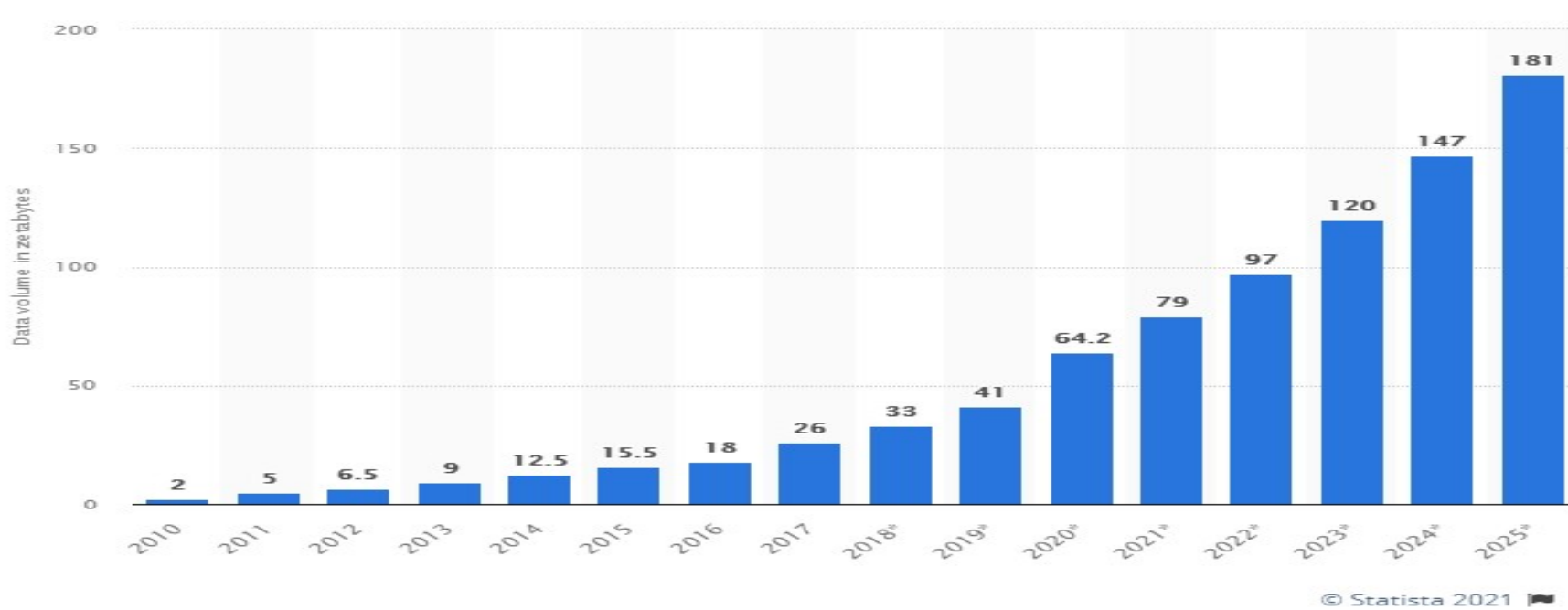


Os 5 v's do Big Data



Crescimento dos dados

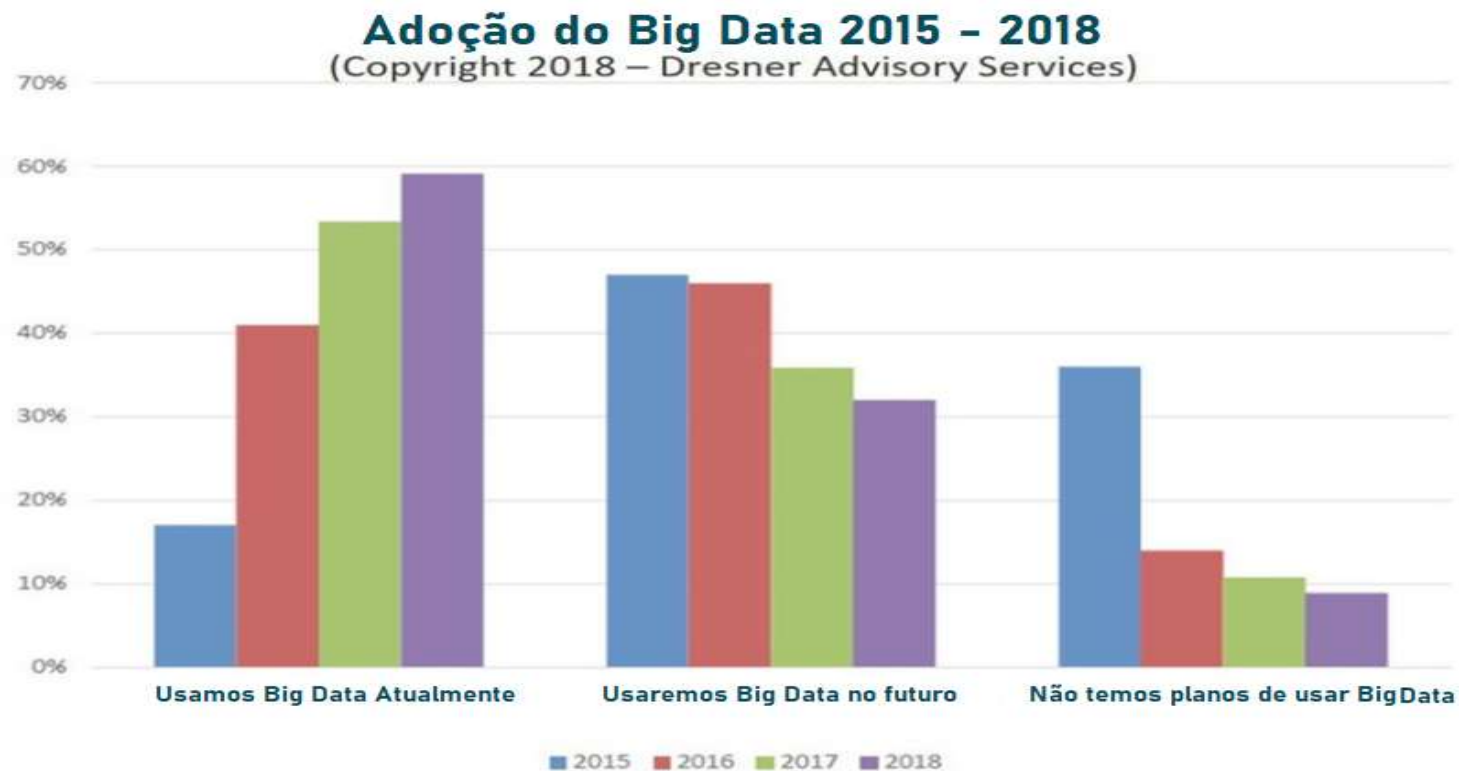
Previsão para até 2025



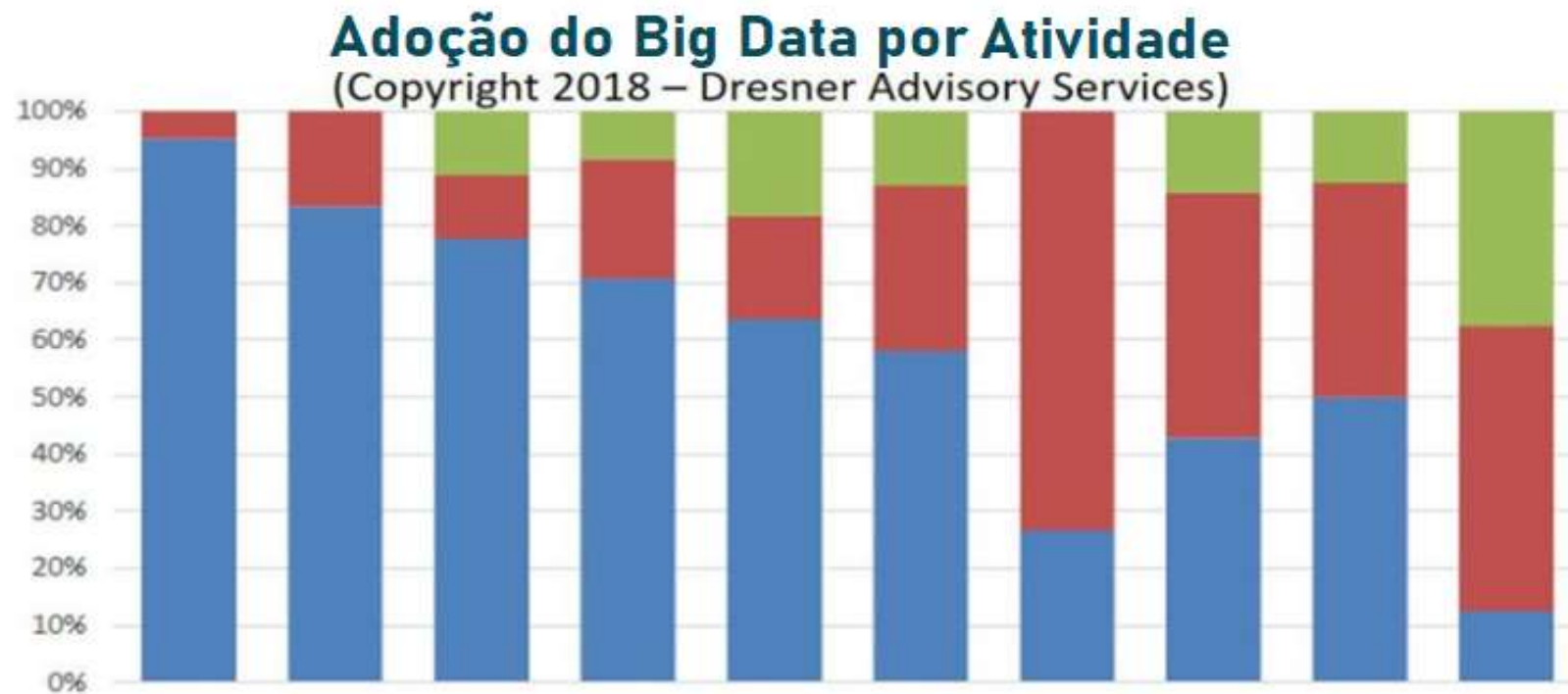
Como estão os dados no mercado

- Google gera 100 PetaBytes de dados por dia.
- Facebook gera 30 ou mais PetaBytes de dados por dia.
- Twitter gera 100 TeraBytes por dia.
- Spotify gera 64 TeraBytes por dia
- eBay gera 100 PetaBytes por dia.
- Até 2007 tínhamos gerado 300 EB de dados, hoje já excedemos a casa dos 4.000 EB,
- Em 2020 geramos mais 50.000 EB ou 50 ZetaBytes.
- O Google é a maior empresa de Big Data do mundo, processando 3,5 bilhões de solicitações diárias,
- gerando e armazenando 10 ExaByte de dados.
- Noventa (90%) de todos os dados do mundo foram produzidos nos últimos 2 anos.

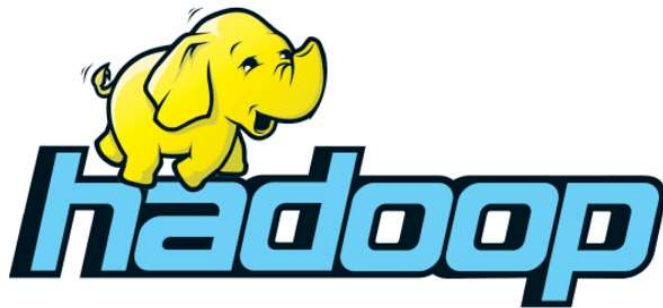
Adoção do Big Data pelas Empresas



Adoção do Big Data pelas Empresas



Haddop



minsait

03

An Indra company

19

Criadores

Doug Cutting



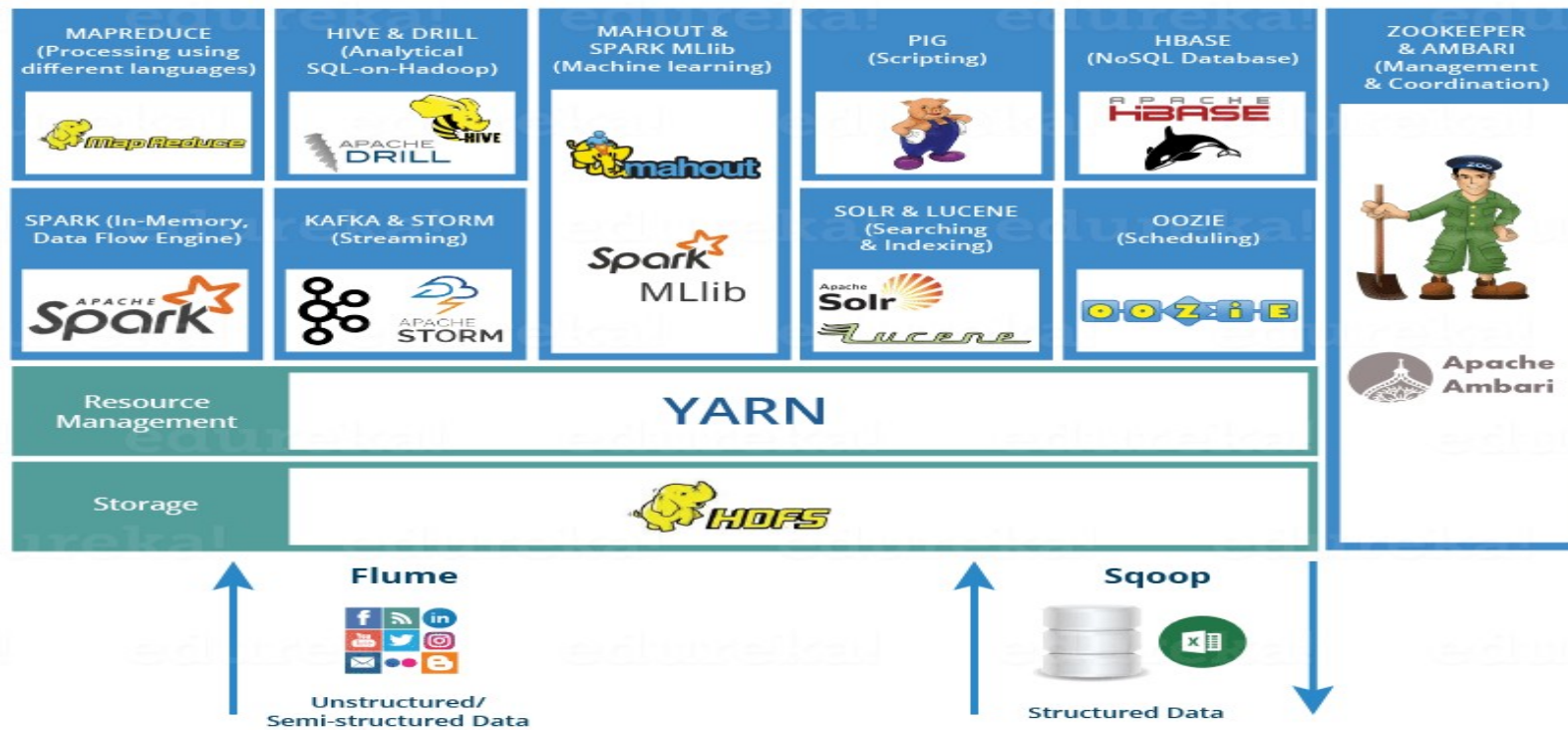
Mike Cafarella



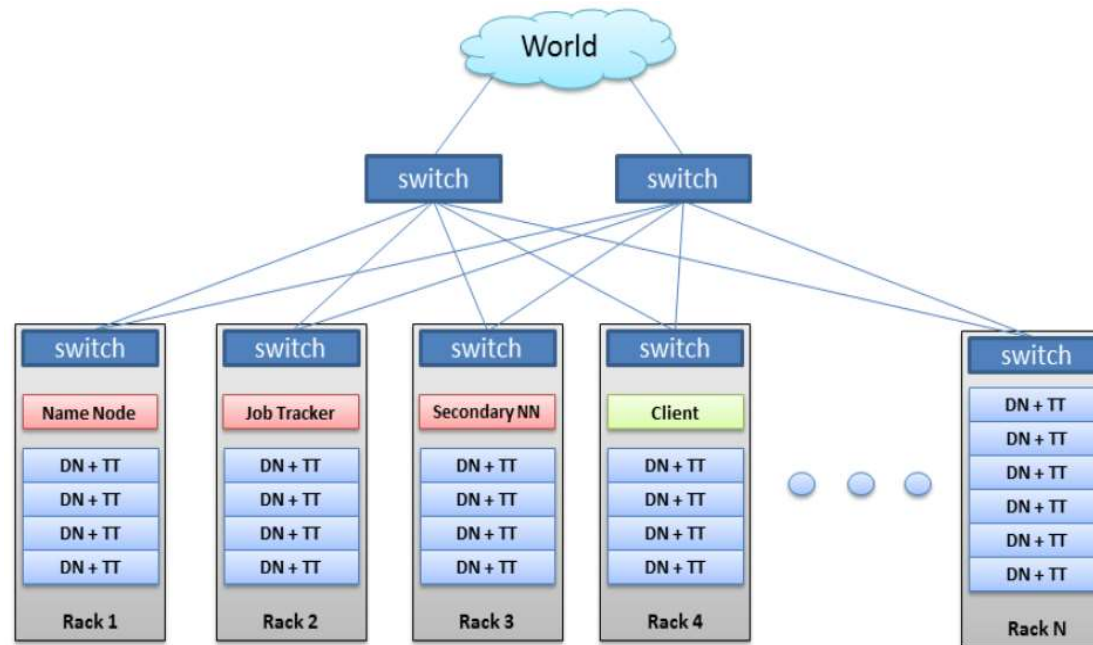
Hadoop o que é?

- Plataforma que fornece infraestrutura econômica e escalável
- Processamento em lote para grandes quantidades de dados
- Armazenamento e Processamento distribuído
- Deu origem ao ecossistema Big Data
- 4 Módulos: HDFS, Mapreduce, Hadoop Common e Yarn

Arquitetura Haddop

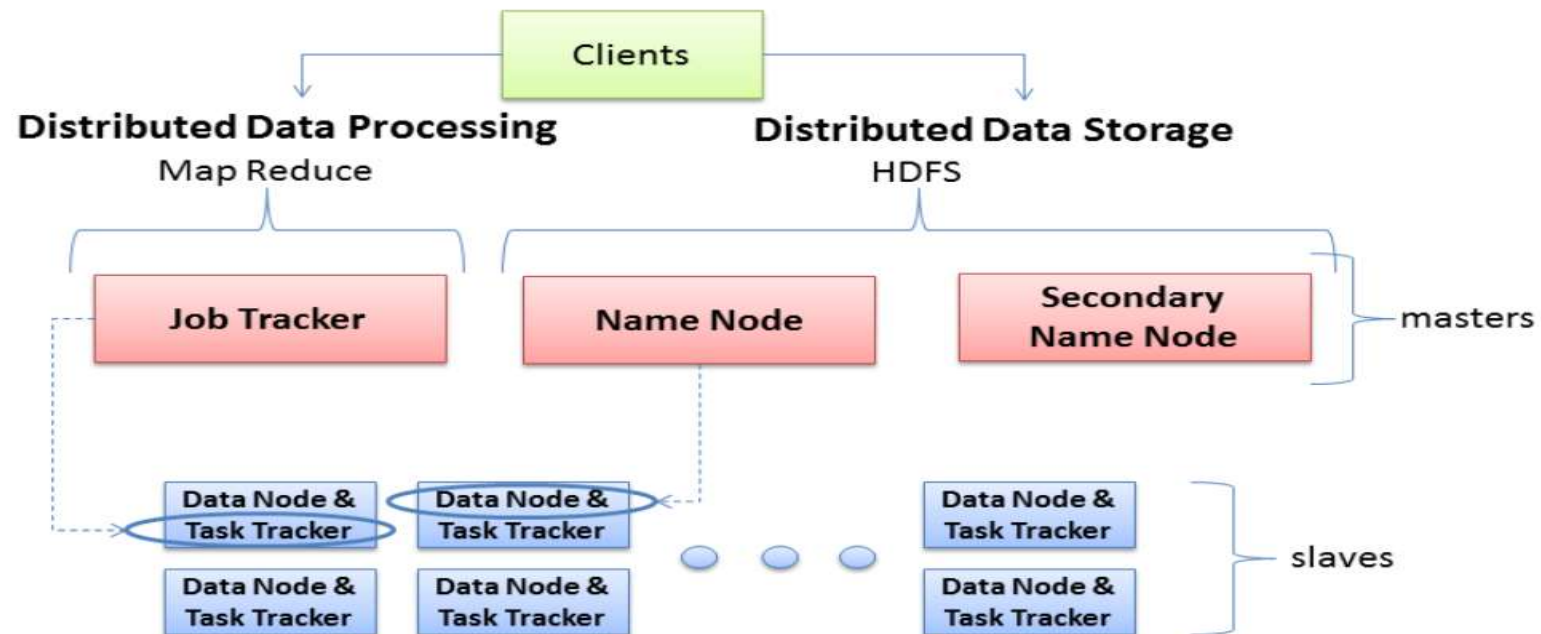


Arquitetura Hadoop



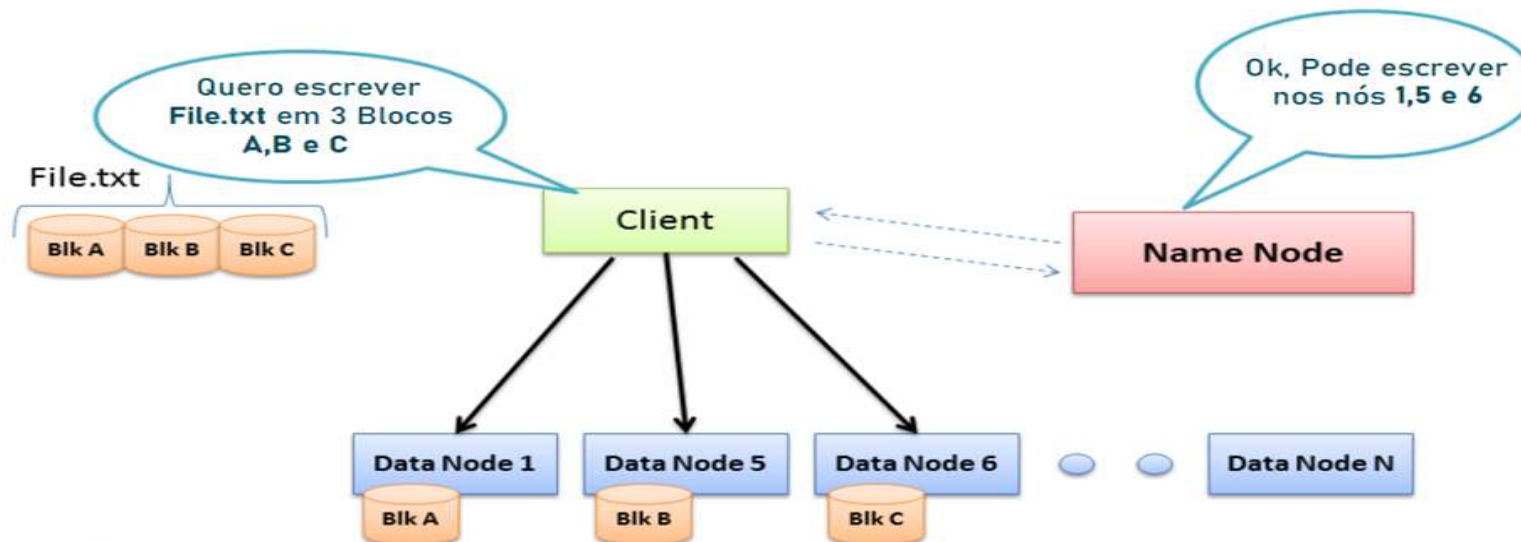
Funcionamento

Arquitetura Hadoop



Funcionamento

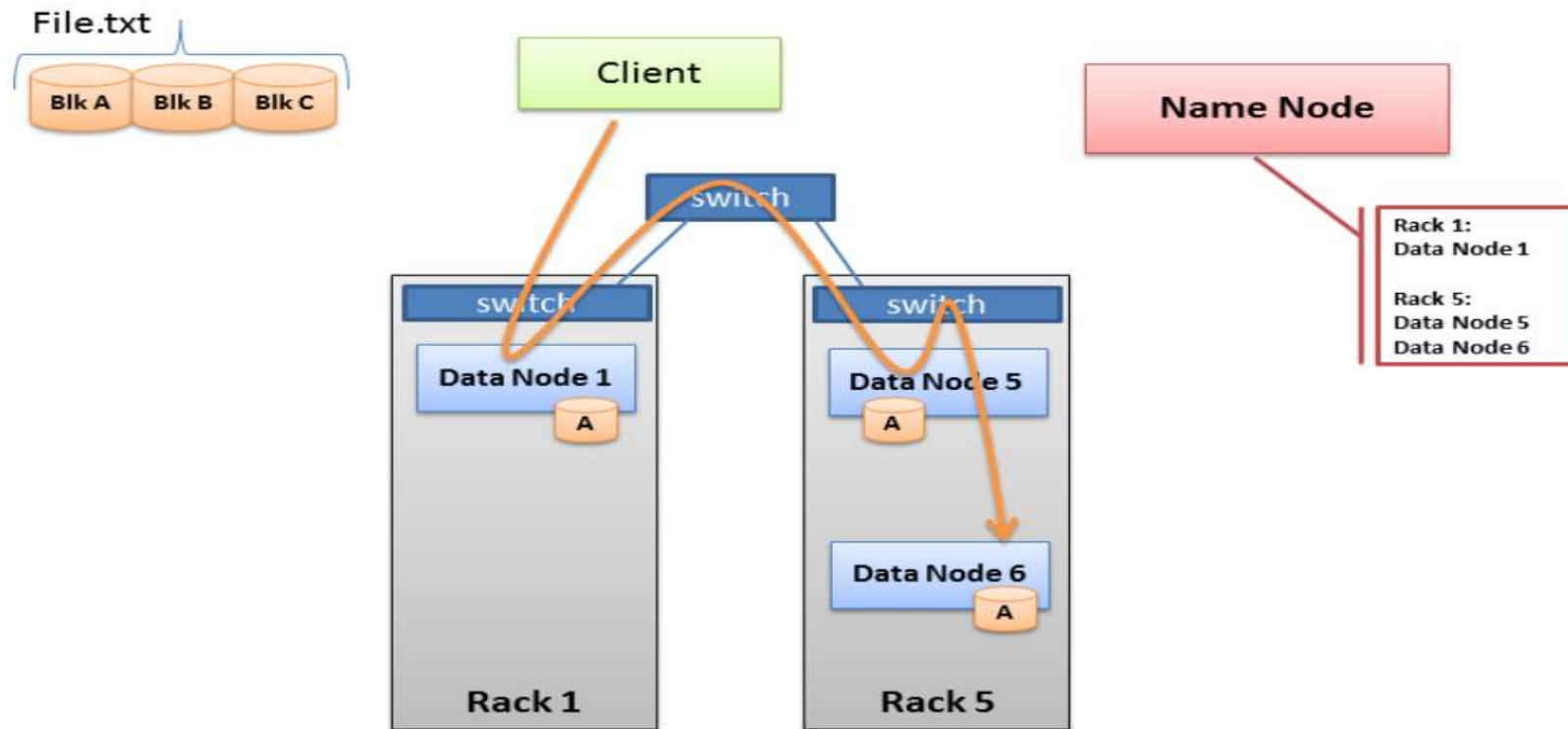
Arquitetura Hadoop



- Cliente consulta Name Node
- Cliente escreve no Data Node
- Data Node replica o bloco
- Ciclo se repete para o próximo bloco

Funcionamento

Arquitetura Hadoop



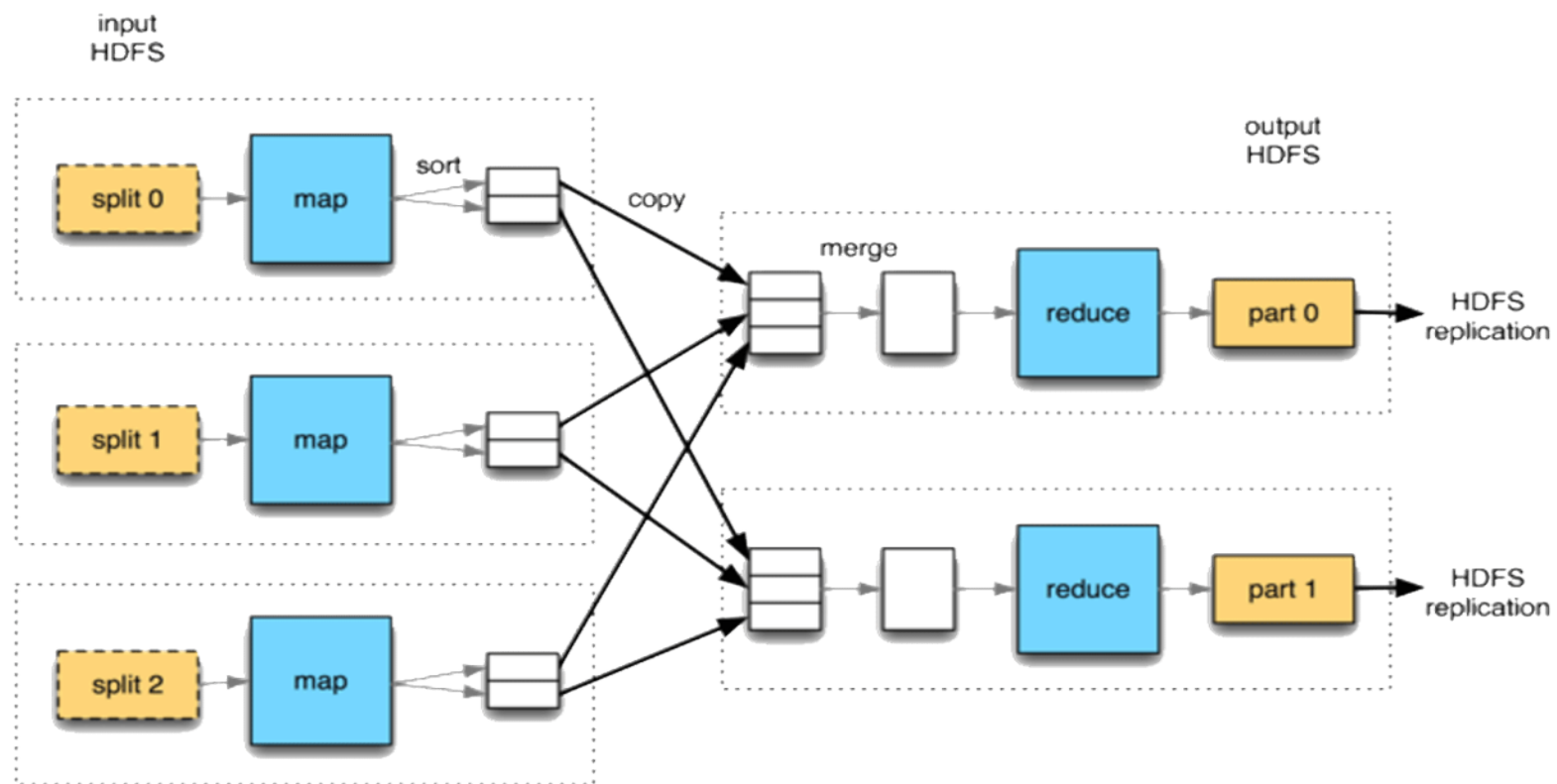
minsoit

An Indra company

26

Funcionamento

Arquitetura Hadoop



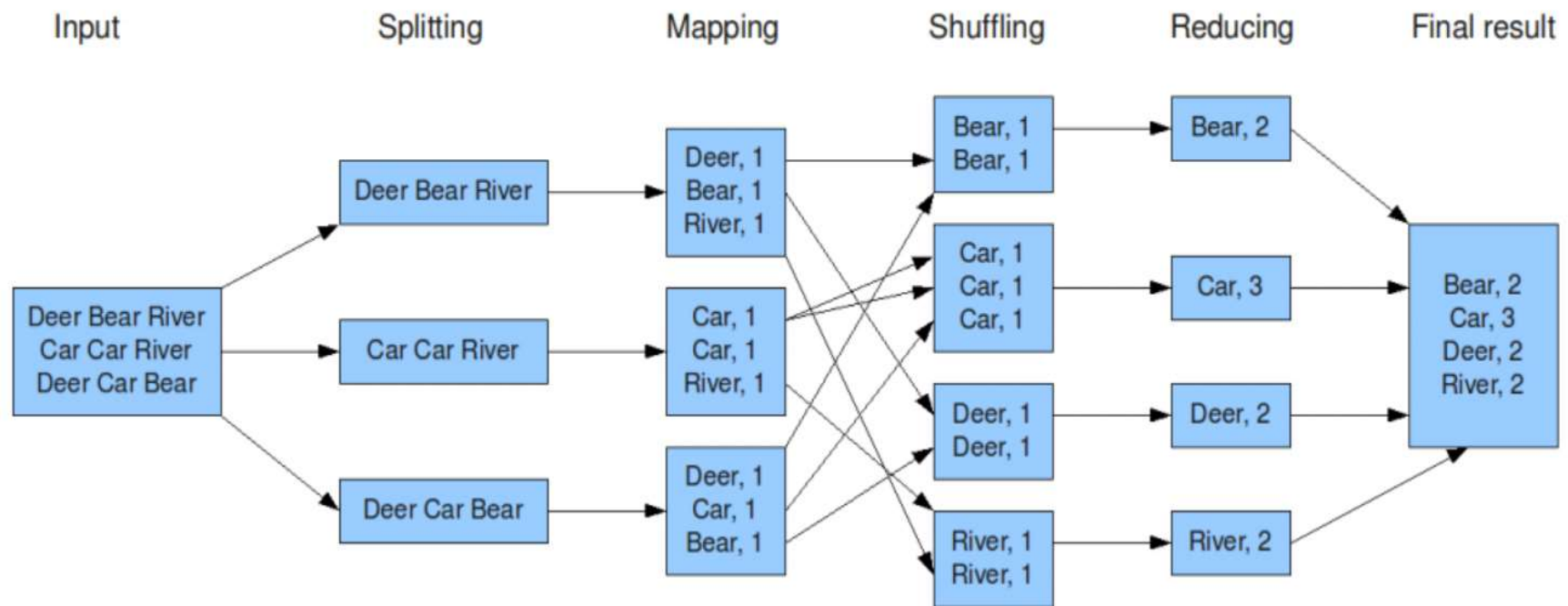
minsoit

An Indra company

27

Funcionamento

Arquitetura Hadoop



minsait

An Indra company

28

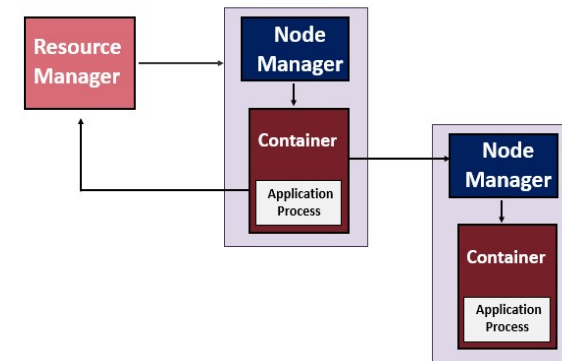
Arquitetura Hadoop - Yarn

Considere o YARN como o cérebro do ecossistema Hadoop.

Executa todas as atividades de processamento, alocando recursos e agendando tarefas.

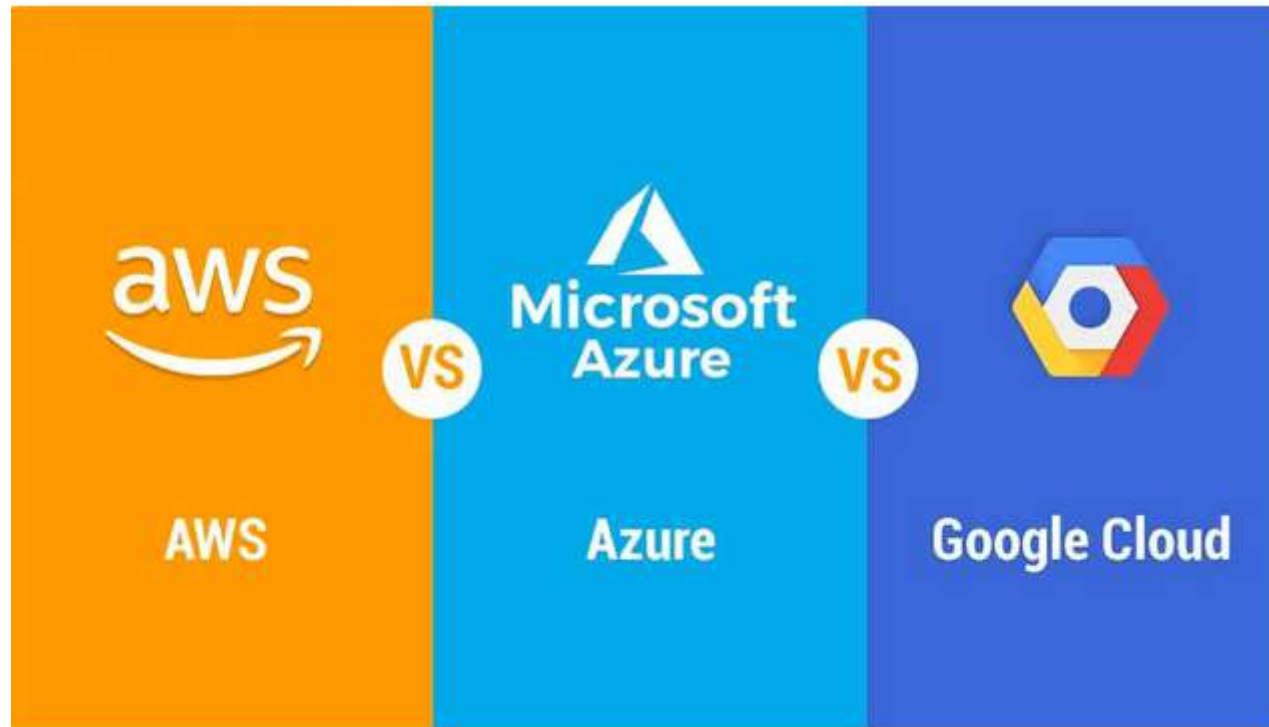
ResourceManager e NodeManager.

- **ResourceManager** é um nó principal ele recebe as solicitações de processamento e, em seguida, passa as partes das solicitações para os NodeManagers correspondentes, onde o processamento real ocorre.
- **NodeManagers** são instalados em cada DataNode. É responsável pela execução da tarefa em cada DataNode único.



Funcionamento

Para onde estamos indo?

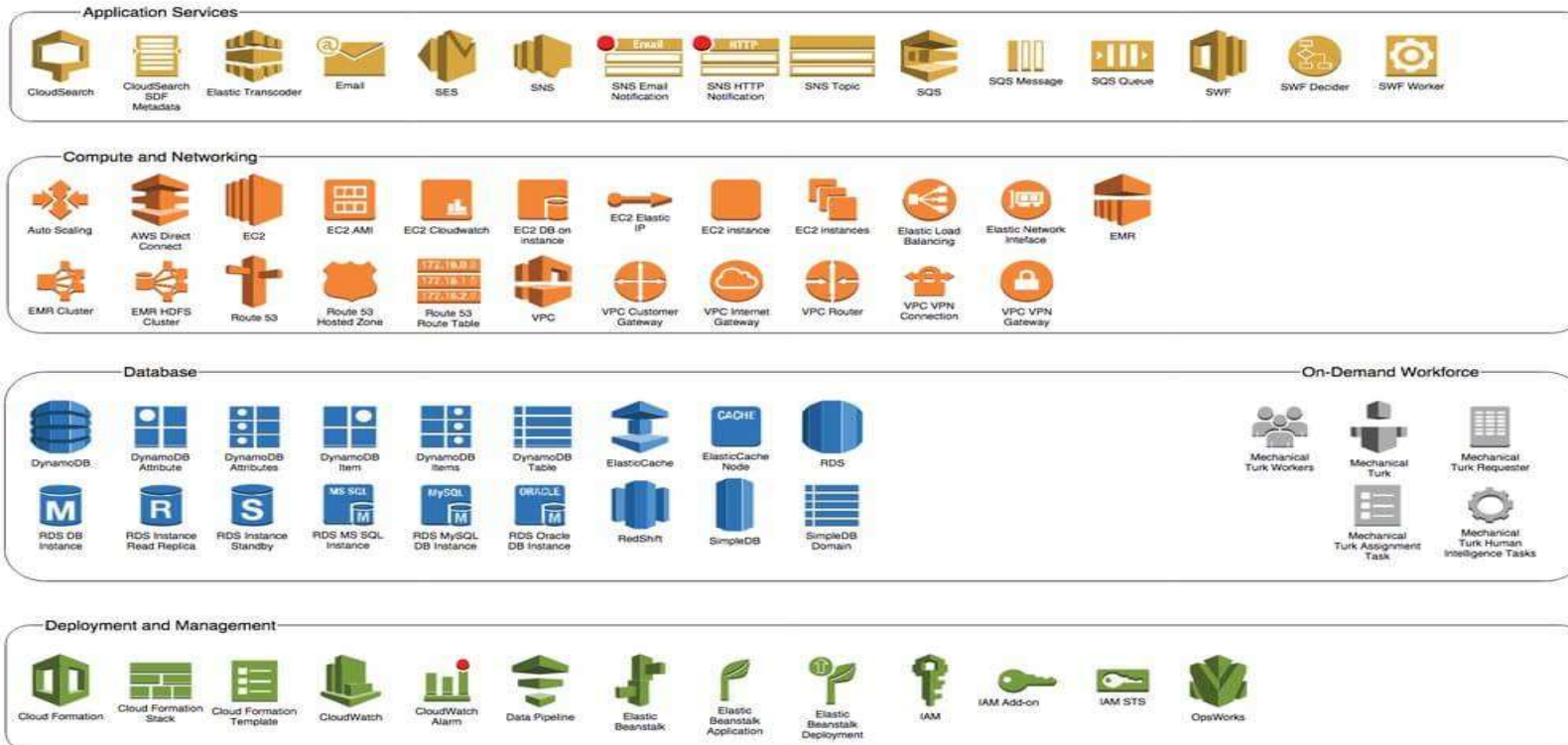


minsait

An Indra company

30



Serviços na AWS



Serviços na AZURE

Infrastructure Services

Compute

-  Windows
-  Linux
-  Containers

Storage

-  BLOB Storage
-  Azure Files
-  Premium Storage

Networking

-  Virtual Network
-  Load Balancer
-  DNS
-  Express Route
-  Traffic Manager
-  VPN Gateway
-  Application Gateway

Platform Services

Compute

-  Cloud Services
-  Service Fabric
-  Batch

Integration

-  Storage Queues
-  Biztalk Services
-  Hybrid Connections
-  Service Bus

Media & CDN

-  Media Services
-  Content Delivery Network (CDN)

App Service

-  Web Apps
-  API Apps
-  API Management
-  Mobile Apps
-  Logic Apps
-  Notification Hubs

Developer Services

-  Visual Studio
-  Azure SDK
-  Team Project
-  Application Insights

Analytics & IoT

-  HDInsight
-  Machine Learning
-  Data Factory
-  Event Hubs
-  Stream Analytics
-  Mobile Engagement

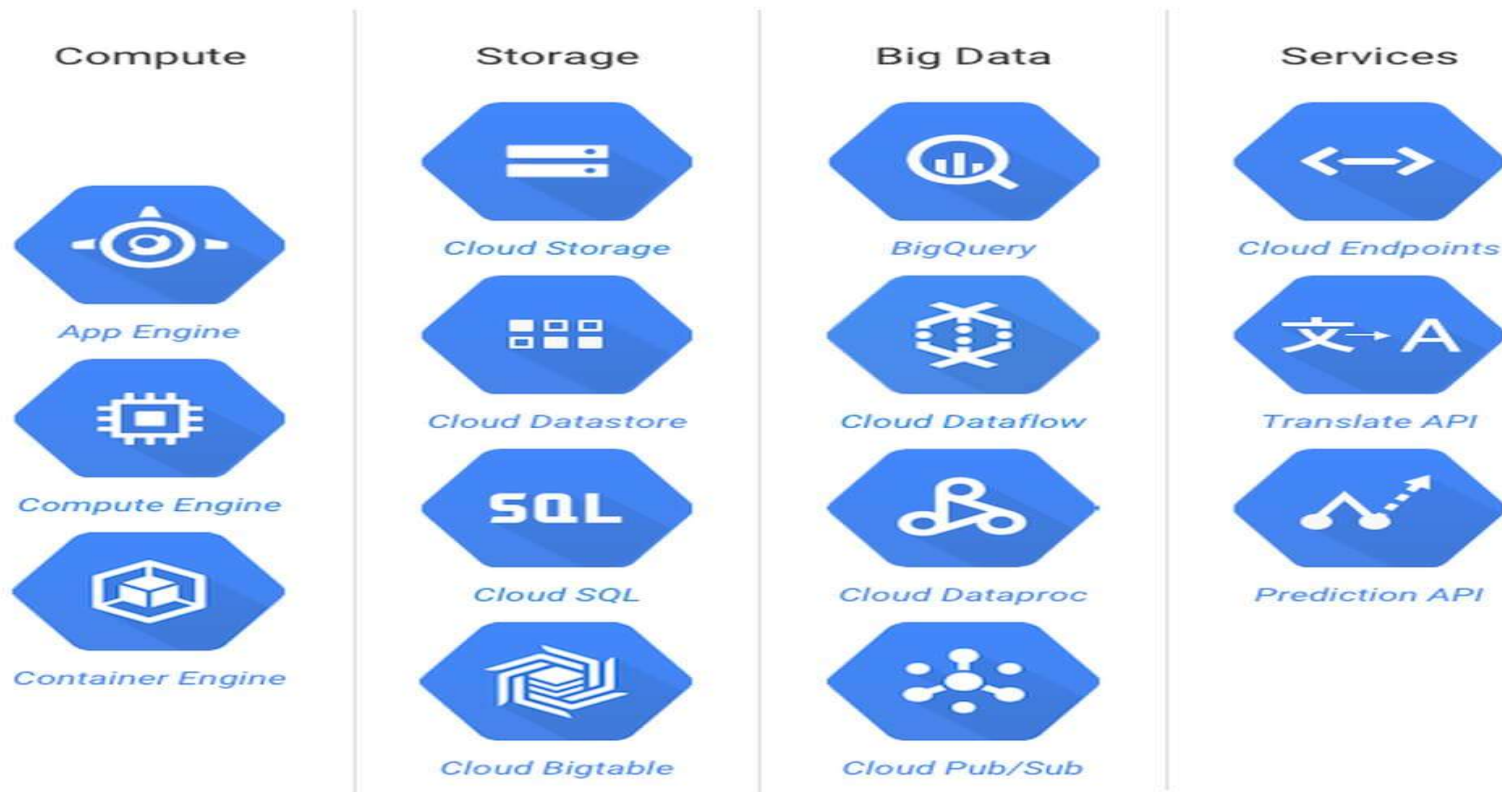
Data

-  SQL Database
-  Redis Cache
-  DocumentDB
-  SQL Data Warehouse
-  Search
-  Tables

Security & Management

-  Portal
-  Active Directory
-  Multi-Factor Authentication
-  Automation
-  Key Vault
-  Store/Marketplace
-  VM Image Gallery & VM Depot

Serviços na GCP



minsait

Mark Making the way forward

An Indra company