DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Seja muito bem-vindo(a),

Ao Curso Data Lake - Design, Projeto e Integração

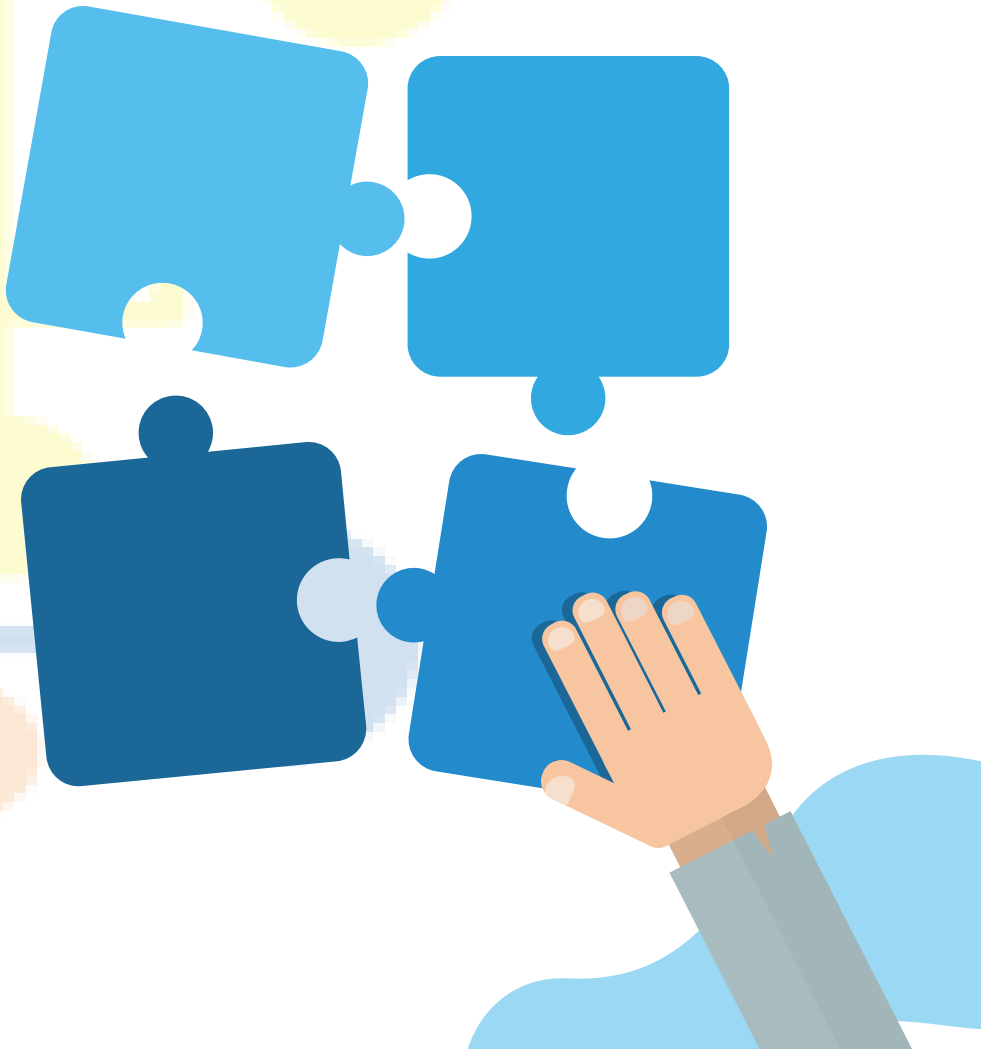
Este é o segundo curso da

Formação Engenheiro de Dados

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

APRESENTAÇÃO

- Introdução
- Instrutores
- Data Science Academy
- Módulos
- Metodologia de Trabalho



Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

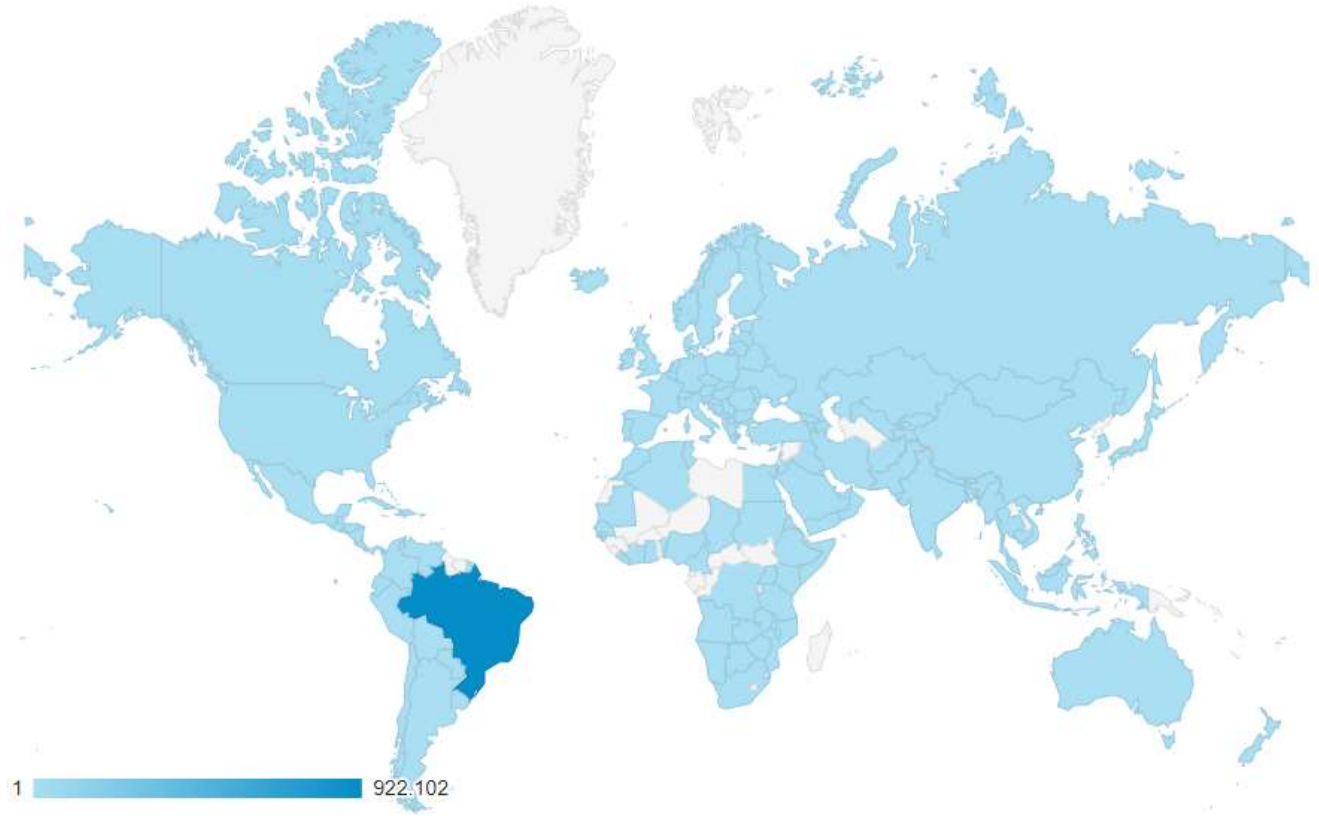
Nossa Academia

A Data Science Academy (DSA) é um portal de ensino online especializado em Big Data, Machine Learning, Inteligência Artificial, Desenvolvimento de Chatbots e tecnologias relacionadas. Nosso objetivo é fornecer aos alunos conteúdo de alto nível por meio do uso de computador, tablet ou smartphone, em qualquer lugar, a qualquer hora, 100% online e 100% em português.

Educação Inovação Futuro

Data Science Academy - Localização

No Brasil e no Mundo



NOSSO OBJETIVO

Que você aprenda de verdade.

Módulos do Curso

Módulos do Curso

1- Introdução

Módulos do Curso

1- Introdução

2- O que são Data Lakes

Módulos do Curso

1- Introdução

2- O que são Data Lakes

3- Data Lake Design

Módulos do Curso

1- Introdução

2- O que são Data Lakes

3- Data Lake Design

4- Data Lake Aquisição de Dados em Batch

Módulos do Curso

1- Introdução

2- O que são Data Lakes

3- Data Lake Design

4- Data Lake Aquisição de Dados em Batch

5- Data Lake Aquisição de Dados em Streaming

Módulos do Curso

6- Data Lake

Camada de

Mensagens

Módulos do Curso

6- Data Lake

Camada de

Mensagens

7- Data Lake

Armazenamento

De Dados

Módulos do Curso

6- Data Lake Camada de Mensagens

7- Data Lake Armazenamento De Dados

8- Data Lake Deploy, Rollout e Boas Práticas

Módulos do Curso

6- Data Lake Camada de Mensagens

7- Data Lake Armazenamento De Dados

8- Data Lake Deploy, Rollout e Boas Práticas

9- Adicionando Camadas ao Data Lake

Módulos do Curso

6- Data Lake Camada de Mensagens

7- Data Lake Armazenamento De Dados

8- Data Lake Deploy, Rollout e Boas Práticas

9- Adicionando Camadas ao Data Lake

10- Avaliação e Certificado

Bônus

Bônus

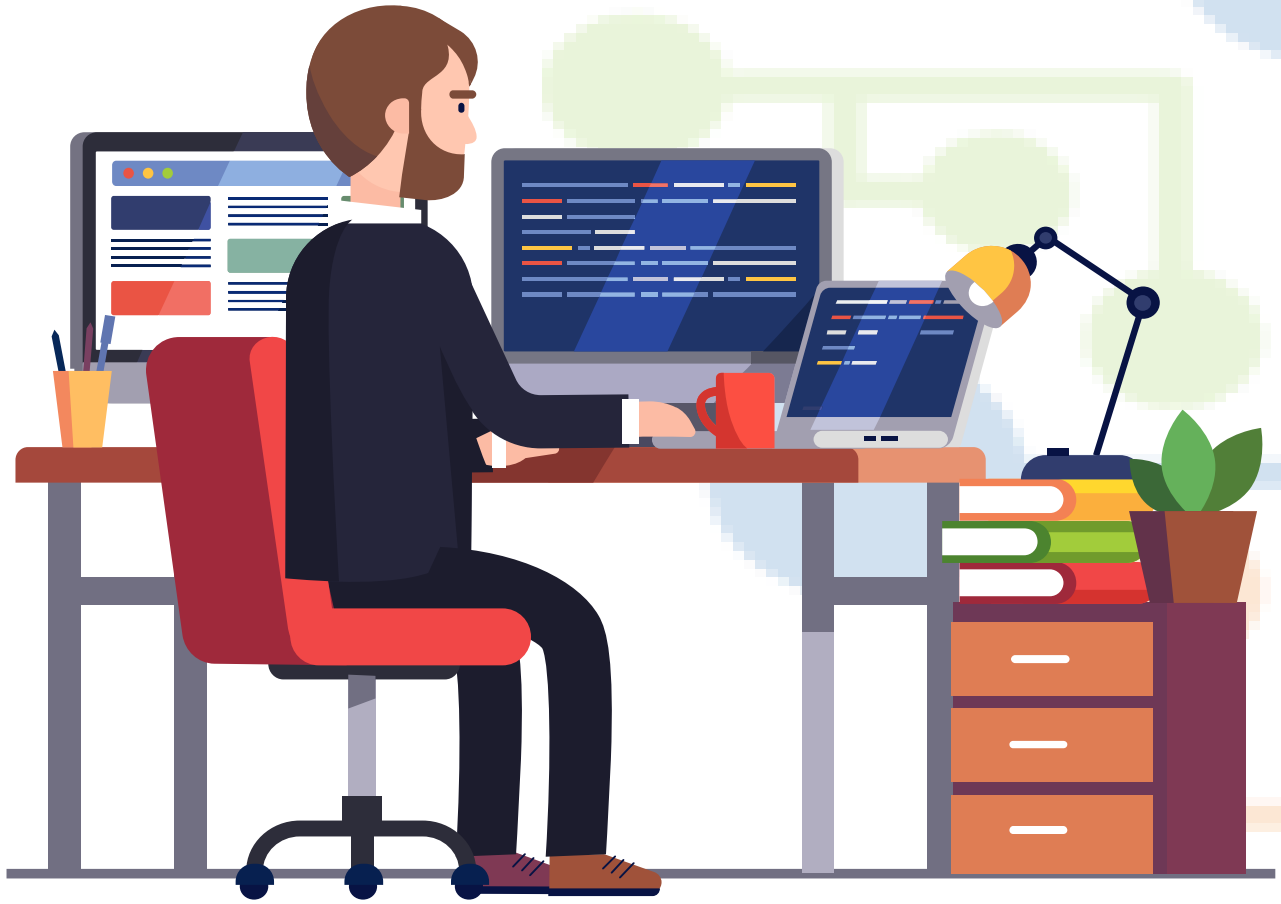
Curso Linux

Bônus

Construindo um Serverless Data Lake

Curso Linux

Curso técnico, voltado para profissionais de tecnologia.



Pré-requisitos

Esperamos que você tenha um bom conhecimento em Sistemas Operacionais (especialmente sistema Linux) e este curso oferece como bônus um curso completo de Introdução ao Sistema Operacional Linux, para ajudar aqueles que não estejam confortáveis com este sistema operacional. Conhecimento de como funciona a Infraestrutura de TI e noções de Business Intelligence.

- Big Data Fundamentos 2.0 e Introdução à Ciência de Dados 2.0
- Sistema Operacional Linux
- Ambiente em Nuvem AWS

O que Esperar Deste Curso?

O que Esperar Deste Curso?

- Teoria e Prática

O que Esperar Deste Curso?

- Teoria e Prática
- Dinamismo

O que Esperar Deste Curso?

- Teoria e Prática
- Dinamismo
- Leitura

O que Esperar Deste Curso?

- Teoria e Prática
- Dinamismo
- Leitura
- Muito Conteúdo

O que Esperar Deste Curso?

- Teoria e Prática
- Dinamismo
- Leitura
- Muito Conteúdo
- Alguns Conceitos Avançados

O que Esperar Deste Curso?

- Teoria e Prática
- Dinamismo
- Leitura
- Muito Conteúdo
- Alguns Conceitos Avançados
- Objetividade

O que NÃO Esperar Deste Curso?

O que NÃO Esperar Deste Curso?

- Segurança e Alta Disponibilidade (estudados no próximo curso da Formação)

O que NÃO Esperar Deste Curso?

- Segurança e Alta Disponibilidade (estudados no próximo curso da Formação)

- Processamento de modelos de ML e IA (estudados na sequência da Formação)

Materiais de Aprendizagem

Aulas em Vídeo
Exposição sobre o conteúdo

Laboratórios
Cenários e Troubleshooting

Pesquisa Adicional
Bibliografia, referências e links úteis ao final de cada capítulo

Quizzes e Exercícios
Quizzes e exercícios para testar seu conhecimento

O que Esperamos de Você?

O que Esperamos de Você?

2 a 4 horas

de dedicação por semana.

O que Esperamos de Você?

2 a 4 horas

de dedicação por semana.

Leitura do Material

E-books e material complementar!

O que Esperamos de Você?

2 a 4 horas

de dedicação por semana.

Leitura do Material

E-books e material complementar!

Interação

Utilize nossas Apps e interaja na rede com outros alunos, no fórum exclusivo e na timeline da Comunidade em nosso site.

O que Esperamos de Você?

2 a 4 horas

de dedicação por semana.

Leitura do Material

E-books e material complementar!

Bibliografia

Leia a bibliografia adicional, acesse os links úteis e realize os quizzes ao final dos capítulos.

Interação

Utilize nossas Apps e interaja na rede com outros alunos, no fórum exclusivo e na timeline da Comunidade em nosso site.

O que Esperamos de Você?

2 a 4 horas

de dedicação por semana.

Divirta-se

Comunique-se, aprenda e divirta-se em nossa Comunidade.

Leitura do Material

E-books e material complementar!

Bibliografia

Leia a bibliografia adicional, acesse os links úteis e realize os quizzes ao final dos capítulos.

Interação

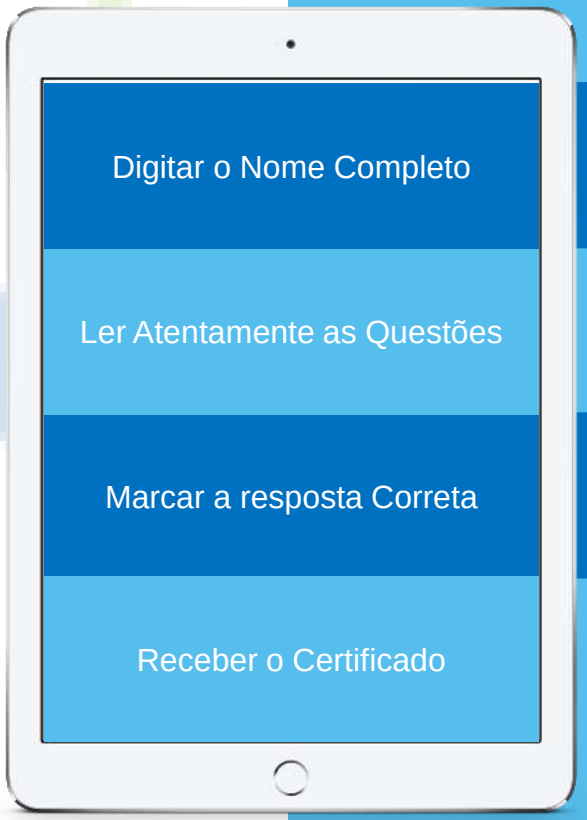
Utilize nossas Apps e interaja na rede com outros alunos, no fórum exclusivo e na timeline da Comunidade em nosso site.

Avaliação Final

3 Tentativas

50 Questões

70 % Aproveitamento

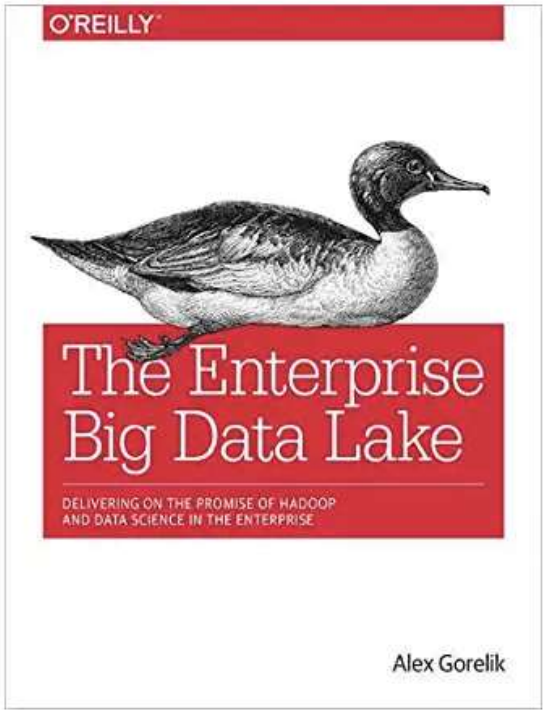


Bibliografia

Blibliografia Sugerida

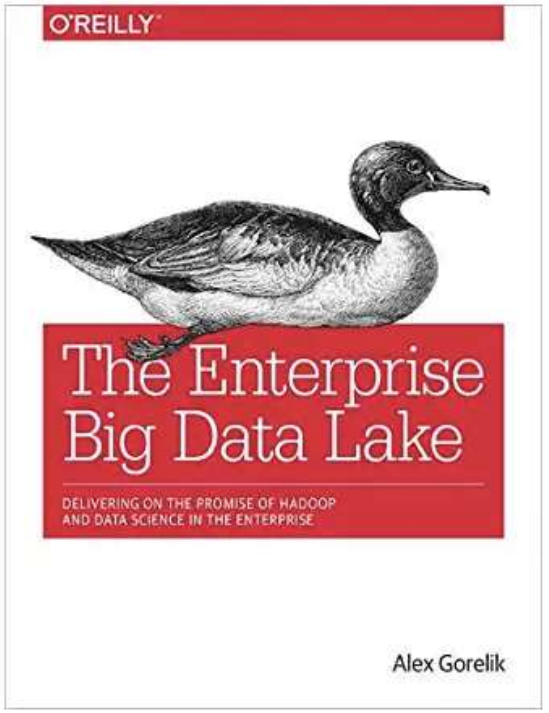
Bibliografia

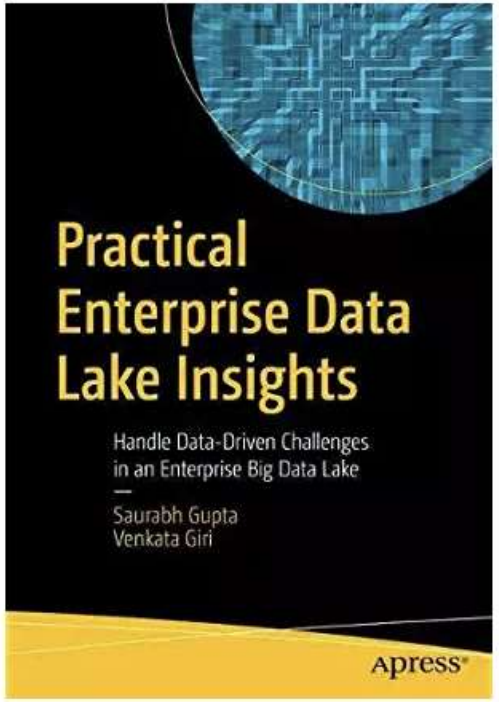
Blibliografia Sugerida



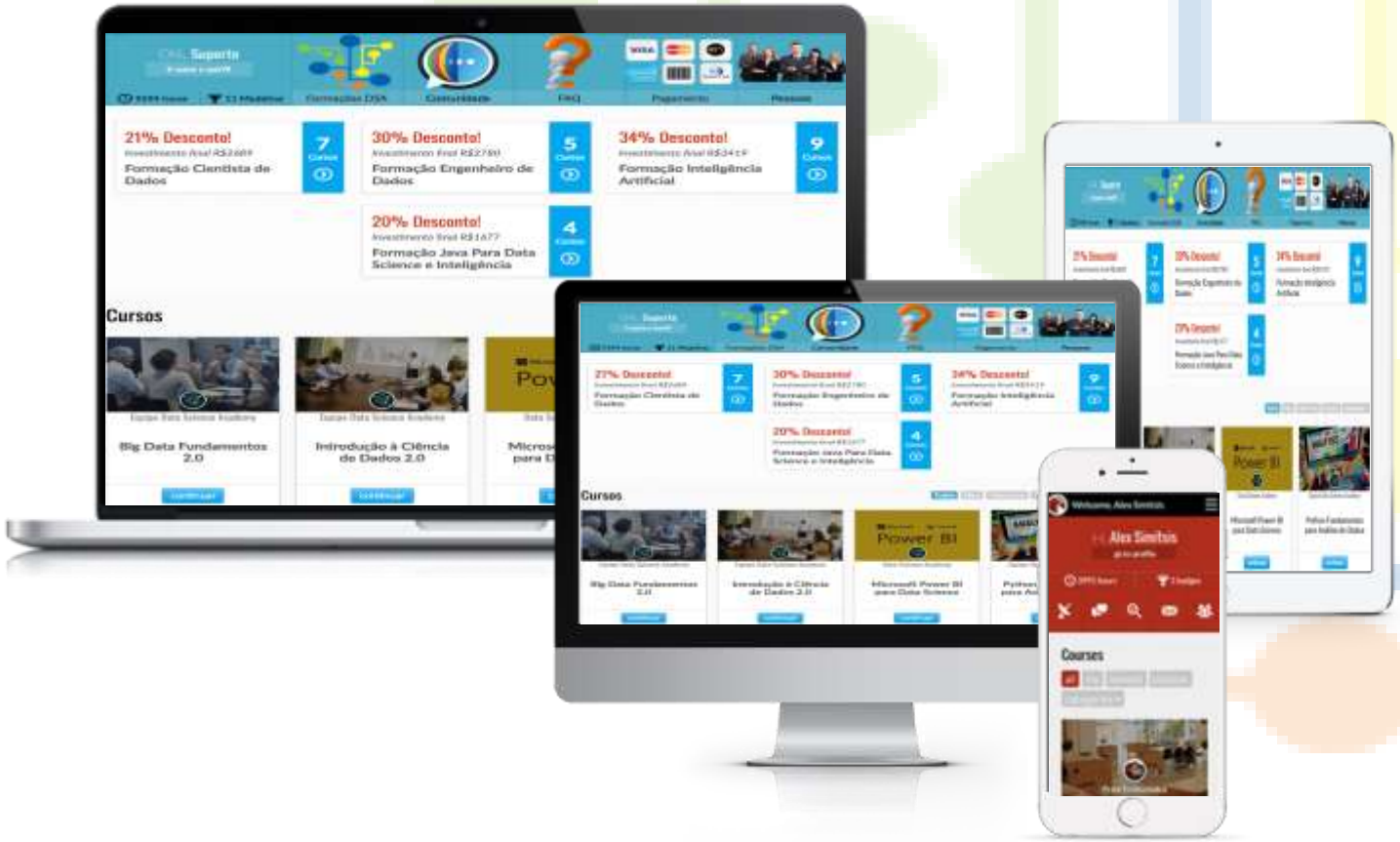
Bibliografia

Blibliografia Sugerida





APPs Gratuitas



iOS

Android

Compartilhe Seu Certificado De Conclusão

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Como Surgiu o Termo Data Lake?

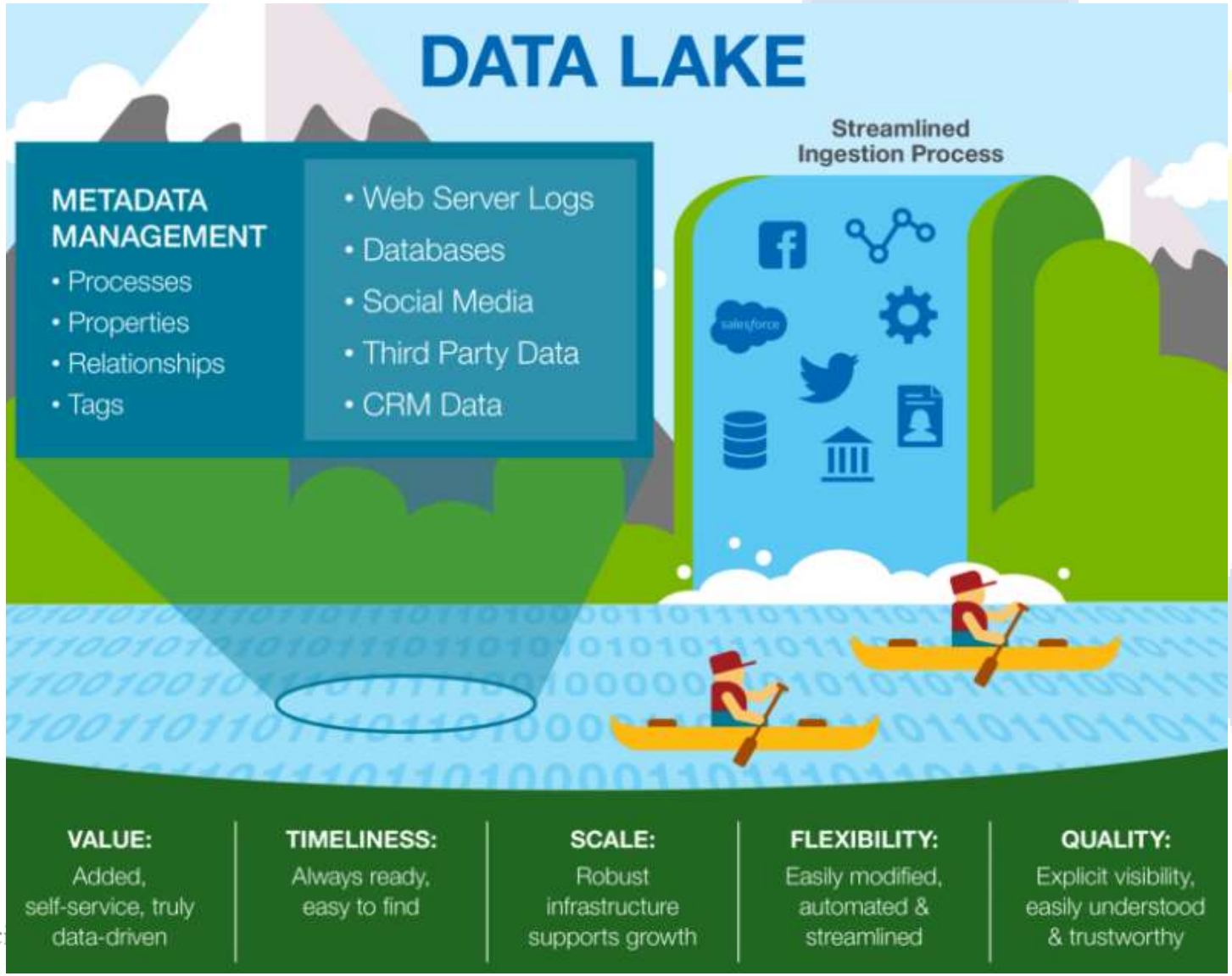
Como Surgiu o Termo Data Lake?

Como Surgiu o Termo Data Lake?

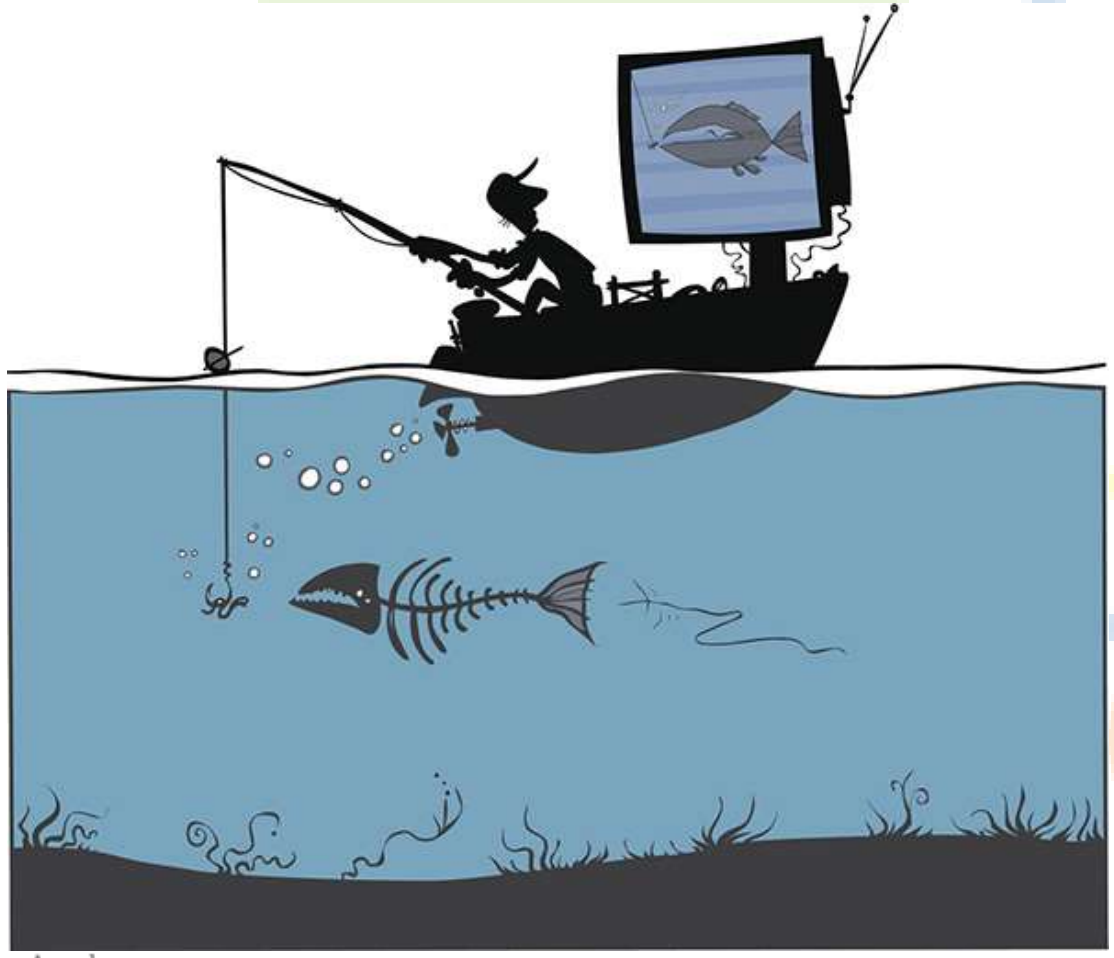
Como Surgiu o Termo Data Lake?

Repositório dentro da empresa, para que todos os dados brutos estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles.

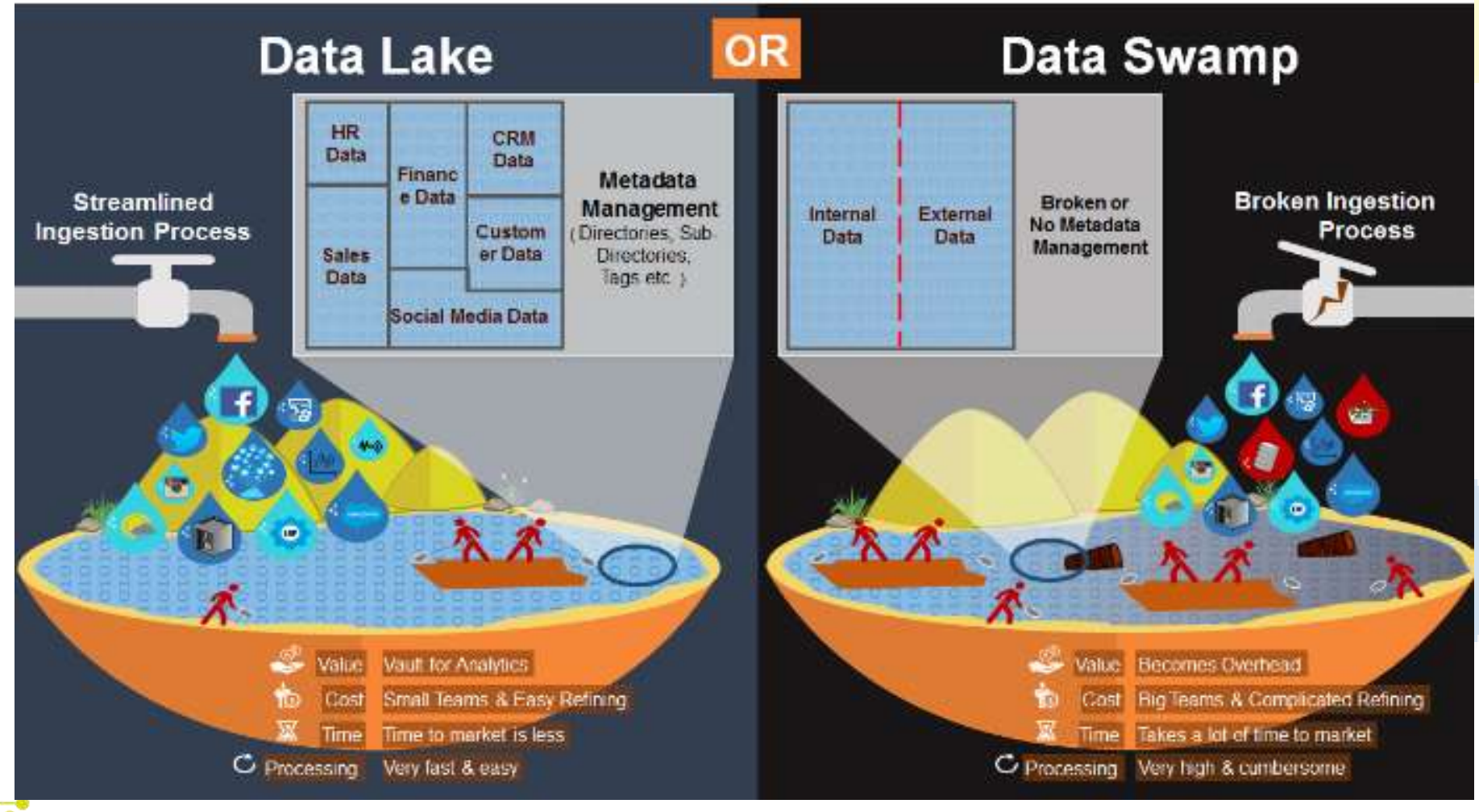
Como Surgiu o Termo Data Lake?



Mas não deixe o Data Lake se tornar um Data Swamp



Data Lake x Data Swamp



Como Vamos Construir o Data Lake?

Como Vamos Construir o Data Lake?

Como Vamos Construir o Data Lake?

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Data Lake
X
Data Warehouse

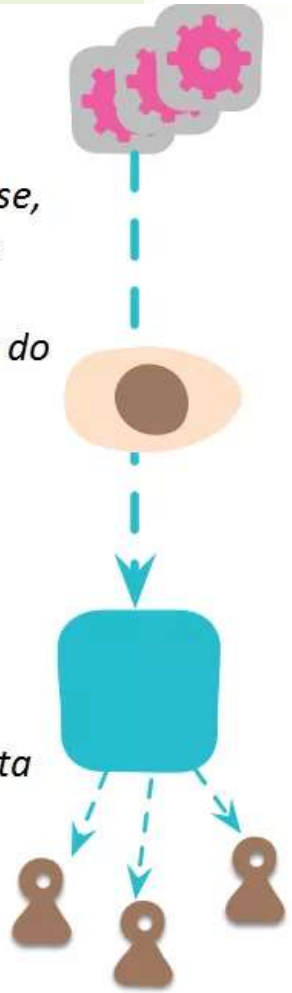
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Data Lake x Data Warehouse

Data Lake x Data Warehouse

Com o Data Warehouse, os dados são limpos e organizados em um único esquema, antes do armazenamento

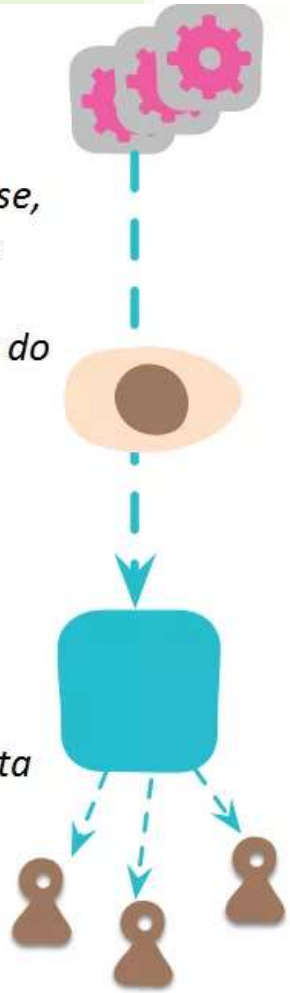
A análise é feita consultando diretamente no Data Warehouse



Data Lake x Data Warehouse

Com o Data Warehouse, os dados são limpos e organizados em um único esquema, antes do armazenamento

A análise é feita consultando diretamente no Data Warehouse



Com o Data Lake, os dados são armazenados em seu formato bruto

Os dados são selecionados e organizados de acordo com a necessidade

Data Lake x Data Warehouse

Isso significa que o Data Warehouse será substituído pelo Data Lake?

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Data Lake x Data Warehouse

Sim e Não

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados		
Processamento		
Armazenamento		
Agilidade		
Segurança		
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento		
Armazenamento		
Agilidade		
Segurança		
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento	Esquema de dados gerado no momento da leitura.	Esquema de dados gerado no momento da escrita.
Armazenamento		
Agilidade		
Segurança		
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento	Esquema de dados gerado no momento da leitura.	Esquema de dados gerado no momento da escrita.
Armazenamento	Criado para ser de baixo custo, independente do volume de dados.	Alto custo para grandes volumes de dados.
Agilidade		
Segurança		
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento	Esquema de dados gerado no momento da leitura.	Esquema de dados gerado no momento da escrita.
Armazenamento	Criado para ser de baixo custo, independente do volume de dados.	Alto custo para grandes volumes de dados.
Agilidade	Bastante ágil. Pode ser configurado e reconfigurado conforme necessário.	Pouco ágil, configuração fixa.
Segurança		
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento	Esquema de dados gerado no momento da leitura.	Esquema de dados gerado no momento da escrita.
Armazenamento	Criado para ser de baixo custo, independente do volume de dados.	Alto custo para grandes volumes de dados.
Agilidade	Bastante ágil. Pode ser configurado e reconfigurado conforme necessário.	Pouco ágil, configuração fixa.
Segurança	Ainda precisa aperfeiçoar o modelo de segurança e acesso aos dados.	Estratégias de segurança bastante maduras.
Usuários		

Data Lake x Data Warehouse

	Data Lake	Data Warehouse
Dados	Estruturados, Semi-estruturados e Não-estruturados. Dados em estado bruto.	Dados estruturados e processados antes da carga no banco de dados.
Processamento	Esquema de dados gerado no momento da leitura.	Esquema de dados gerado no momento da escrita.
Armazenamento	Criado para ser de baixo custo, independente do volume de dados.	Alto custo para grandes volumes de dados.
Agilidade	Bastante ágil. Pode ser configurado e reconfigurado conforme necessário.	Pouco ágil, configuração fixa.
Segurança	Ainda precisa aperfeiçoar o modelo de segurança e acesso aos dados.	Estratégias de segurança bastante maduras.
Usuários	Cientistas e Analistas de Dados	Analistas de Negócio, BI e Cientistas de Dados.

Muito Obrigado.

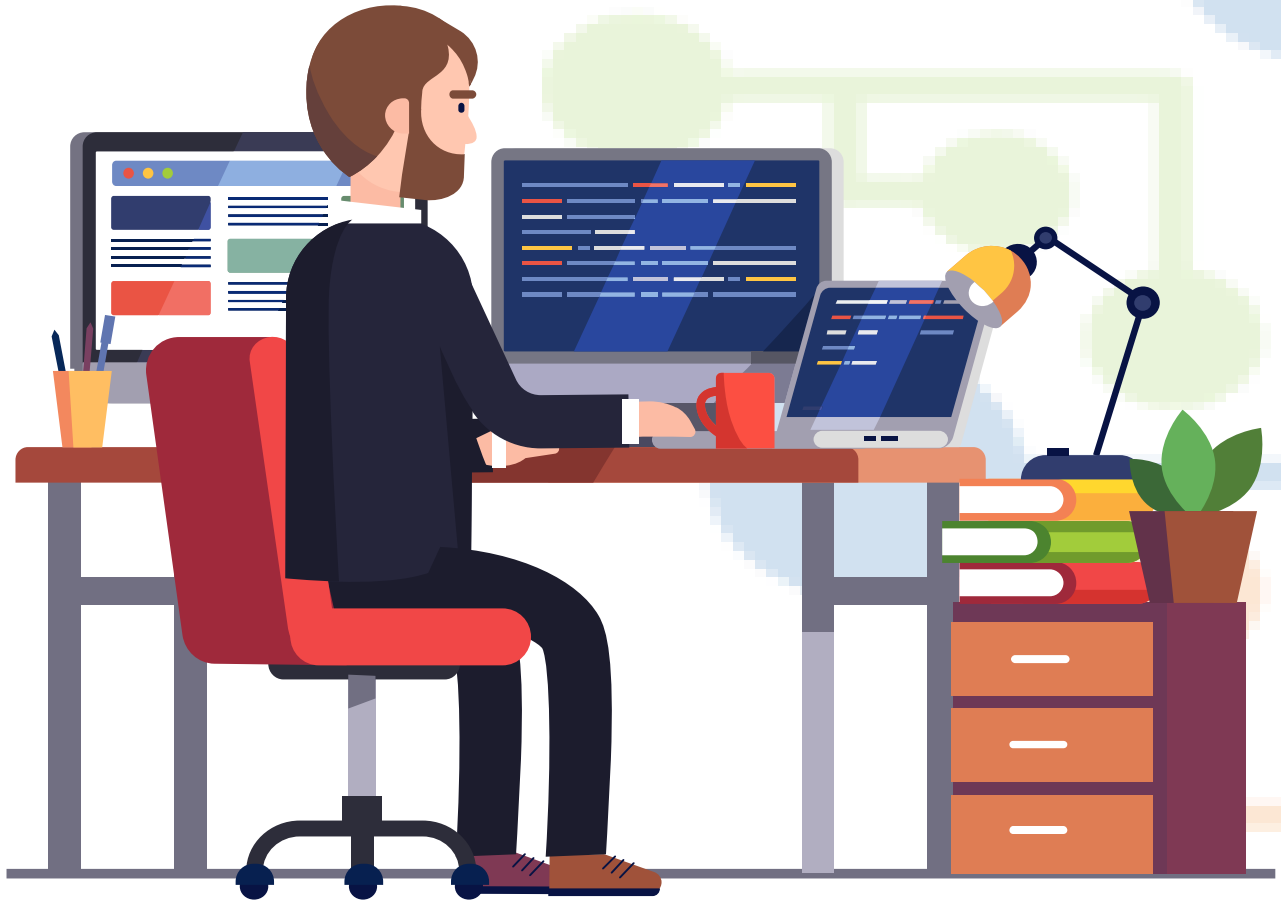
É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

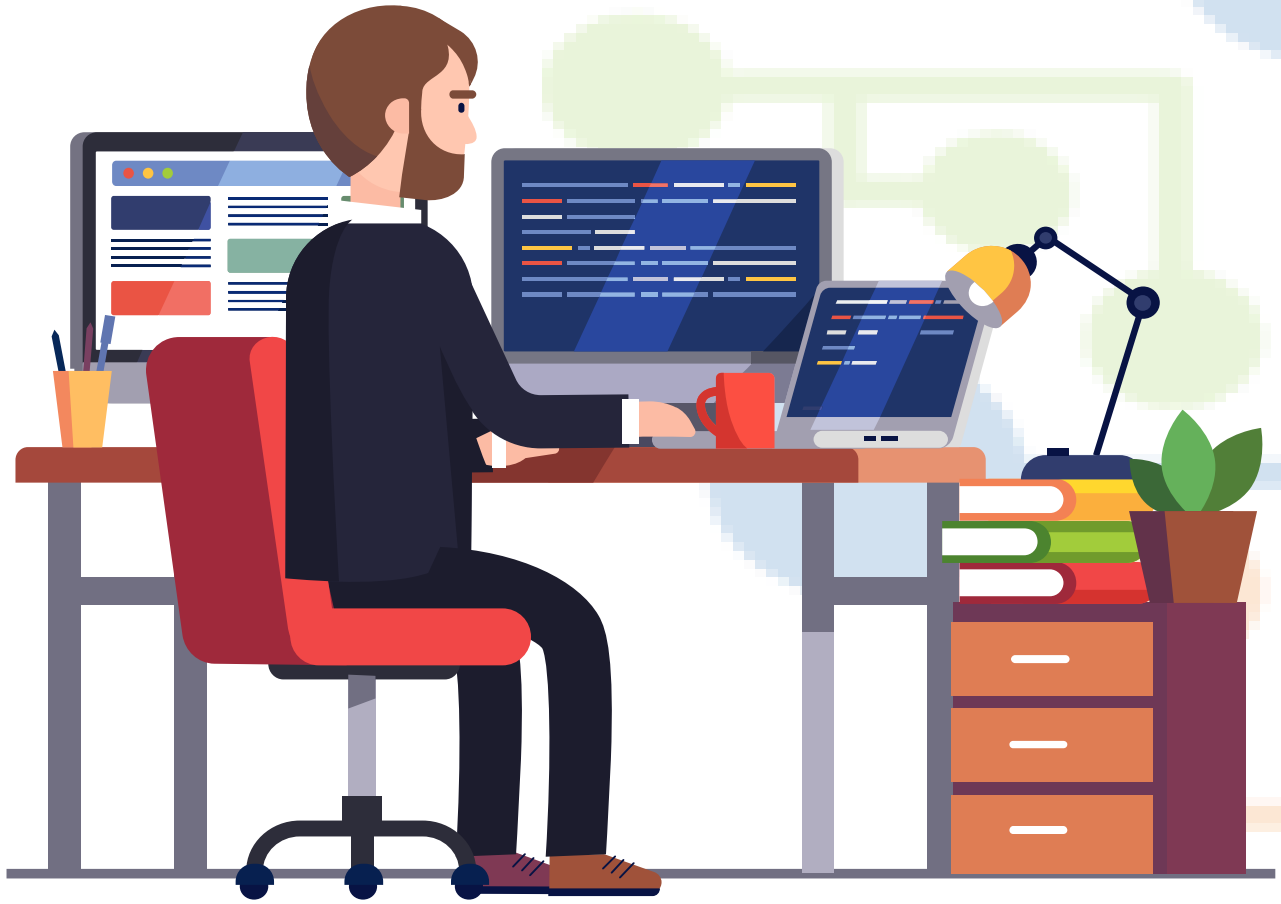
DESIGN, PROJETO E INTEGRAÇÃO

O Que São Dados Corporativos?

O Que São Dados Corporativos?



O Que São Dados Corporativos?



Dados corporativos significam todos e quaisquer dados mantidos por qualquer uma das empresas, incluindo, mas não se limitando a, dados relacionados a suas finanças, impostos, funcionários, clientes, fornecedores e produtos ou serviços.

Construir a Infraestutura Certa é Fundamental

Construir a Infraestutura Certa é Fundamental

Portanto, ela deve ser:

- ✓ Ágil
- ✓ Segura
- ✓ Compatível
- ✓ Gerenciável

Construir a Infraestutura Certa é Fundamental

Portanto, ela deve ser:

- ✓ Ágil
- ✓ Segura
- ✓ Compatível
- ✓ Gerenciável

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Dados

Os dados se tornaram a moeda da economia moderna. Um estudo recente projeta o volume global de dados para crescer de cerca de 0,8 zettabytes (ZB) em 2009 para mais de 35 ZB em 2020, a maior parte gerada nos últimos dois anos e mantida pelo setor corporativo.

Dados corporativos, quando analisados, revelam grandes respostas.

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Dados Estruturados
x
Dados Não Estruturados

Dados Estruturados x Dados Não Estruturados

Dados Estruturados x Dados Não Estruturados

Características	Data Warehouse	Big Data
dados	estruturados	semi-estruturados e não estruturados
volume	grande	escala web
geração	transacional	comportamental

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Infraestrutura em Nuvem
x
On-premise

Infraestrutura em Nuvem x On-premise

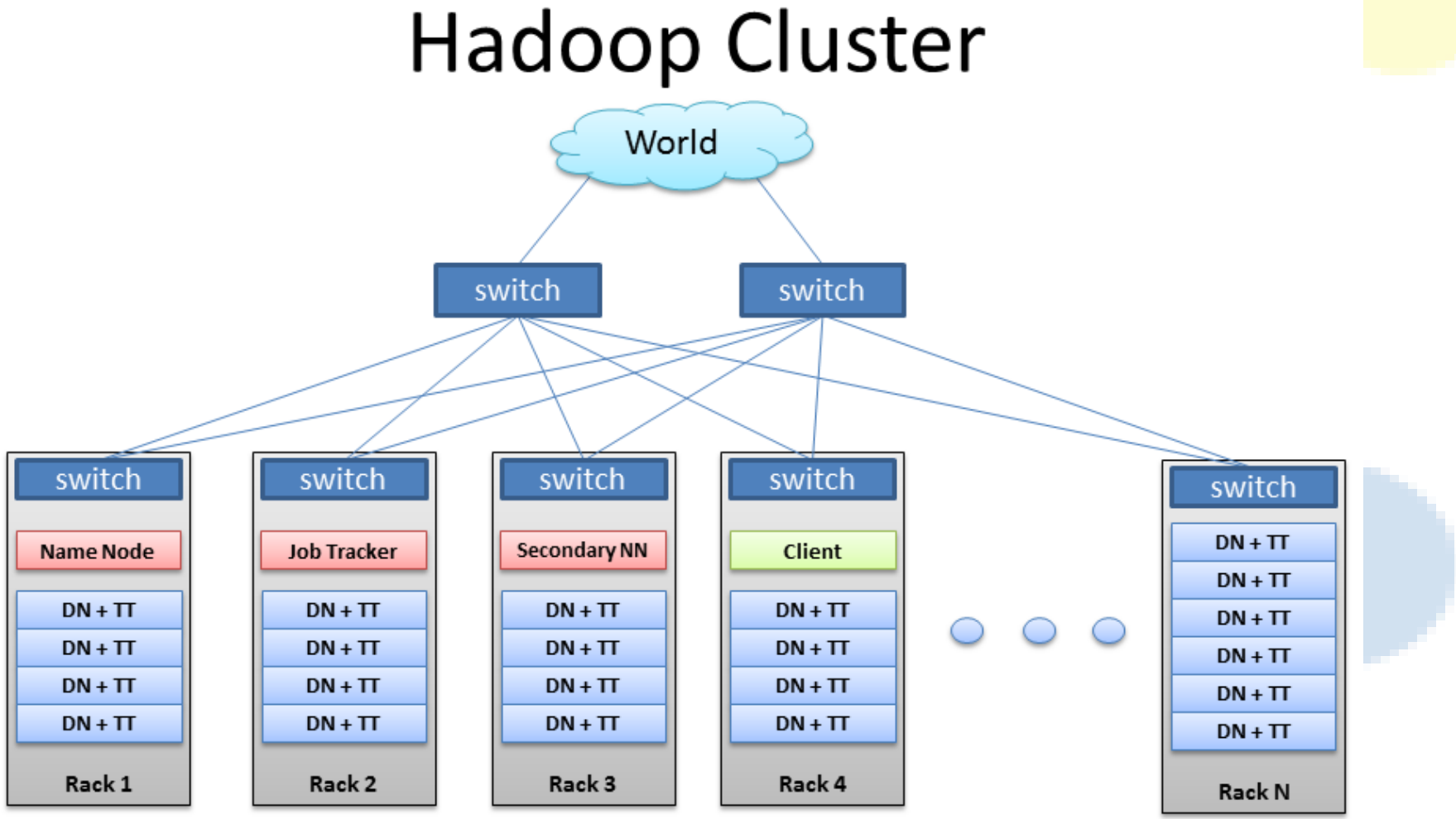
O custo de se produzir um cluster pode ser alto e um Data Lake requer a utilização de um cluster para mostrar o seu valor. Isso nos leva a questão:

Infraestrutura em Nuvem x On-premise

O custo de se produzir um cluster pode ser alto e um Data Lake requer a utilização de um cluster para mostrar o seu valor. Isso nos leva a questão:

Infraestrutura em nuvem ou on-premise?

Infraestrutura em Nuvem x On-premise



Muito Obrigado.

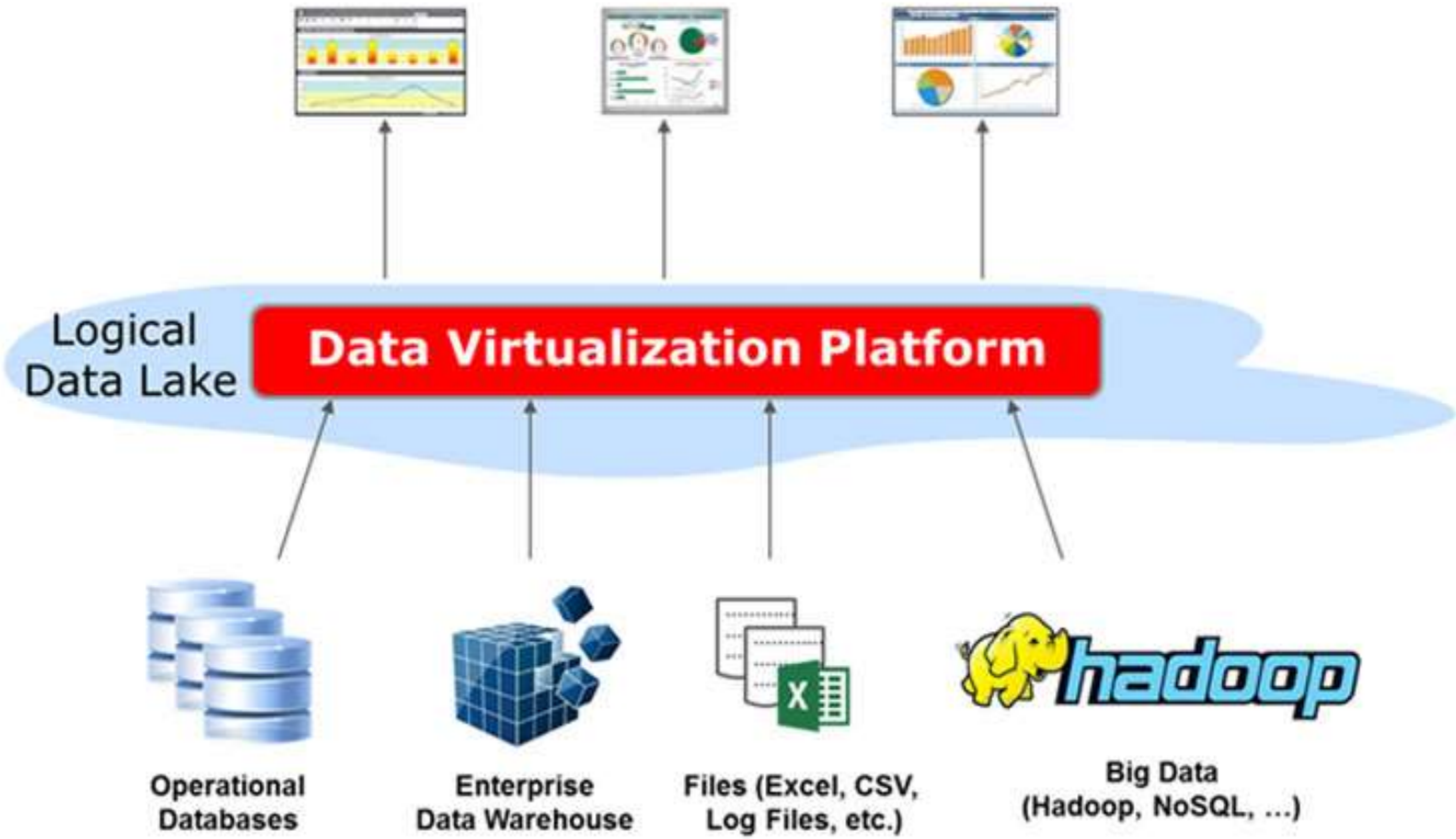
É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

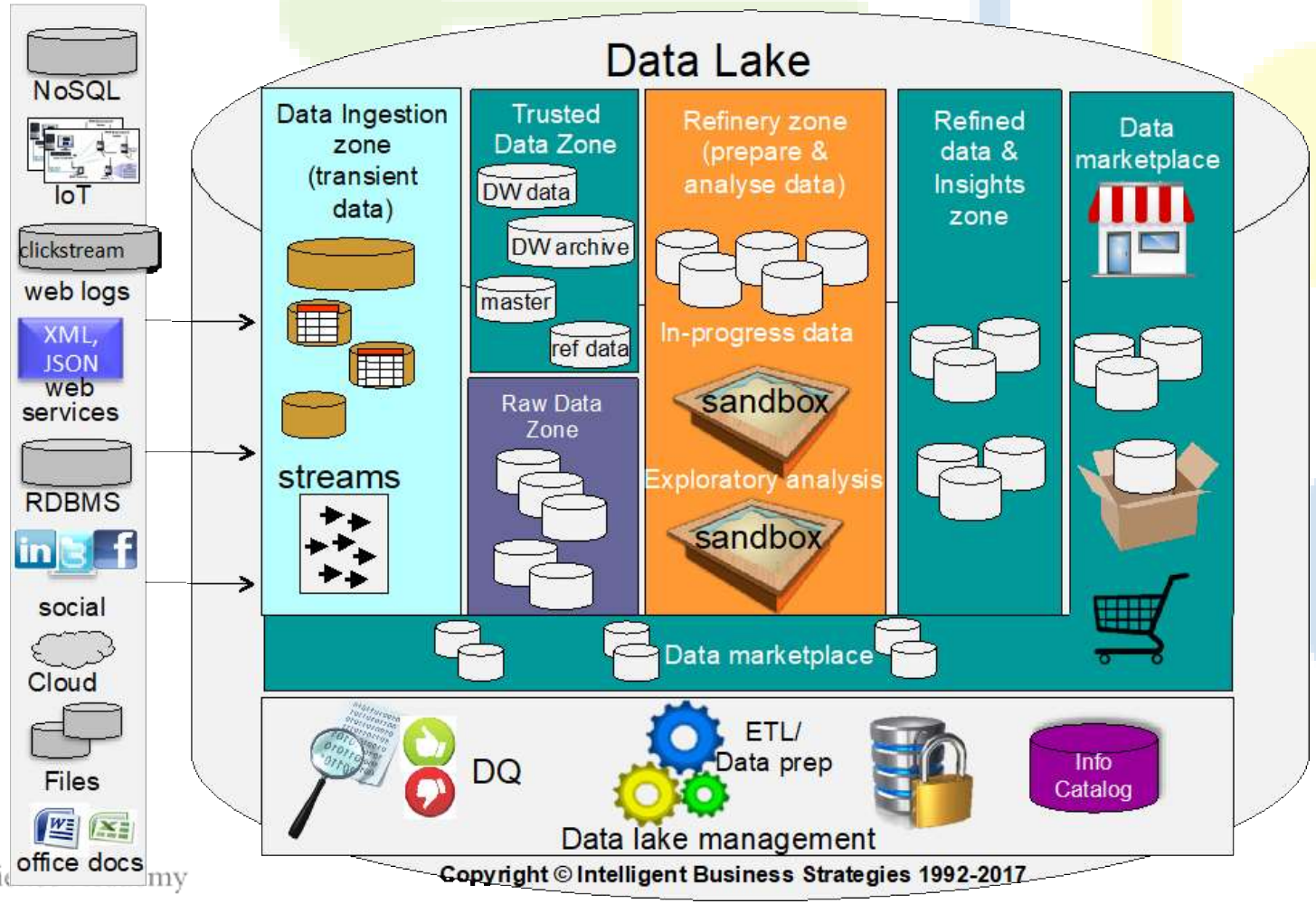
DESIGN, PROJETO E INTEGRAÇÃO

Logical Data Lake

Logical Data Lake



Logical Data Lake



Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

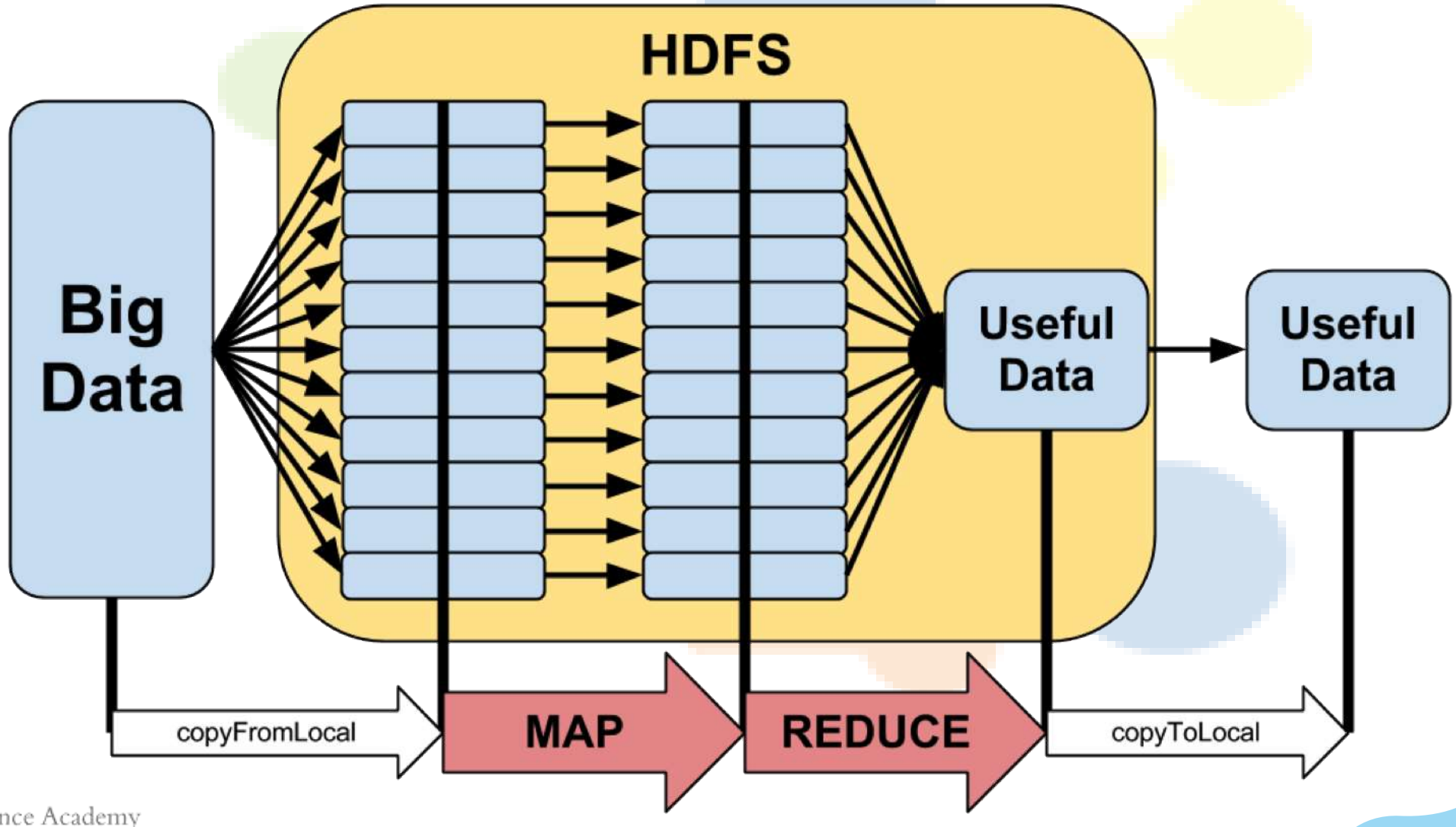
DESIGN, PROJETO E INTEGRAÇÃO

Apache Hadoop e Big Data Stack

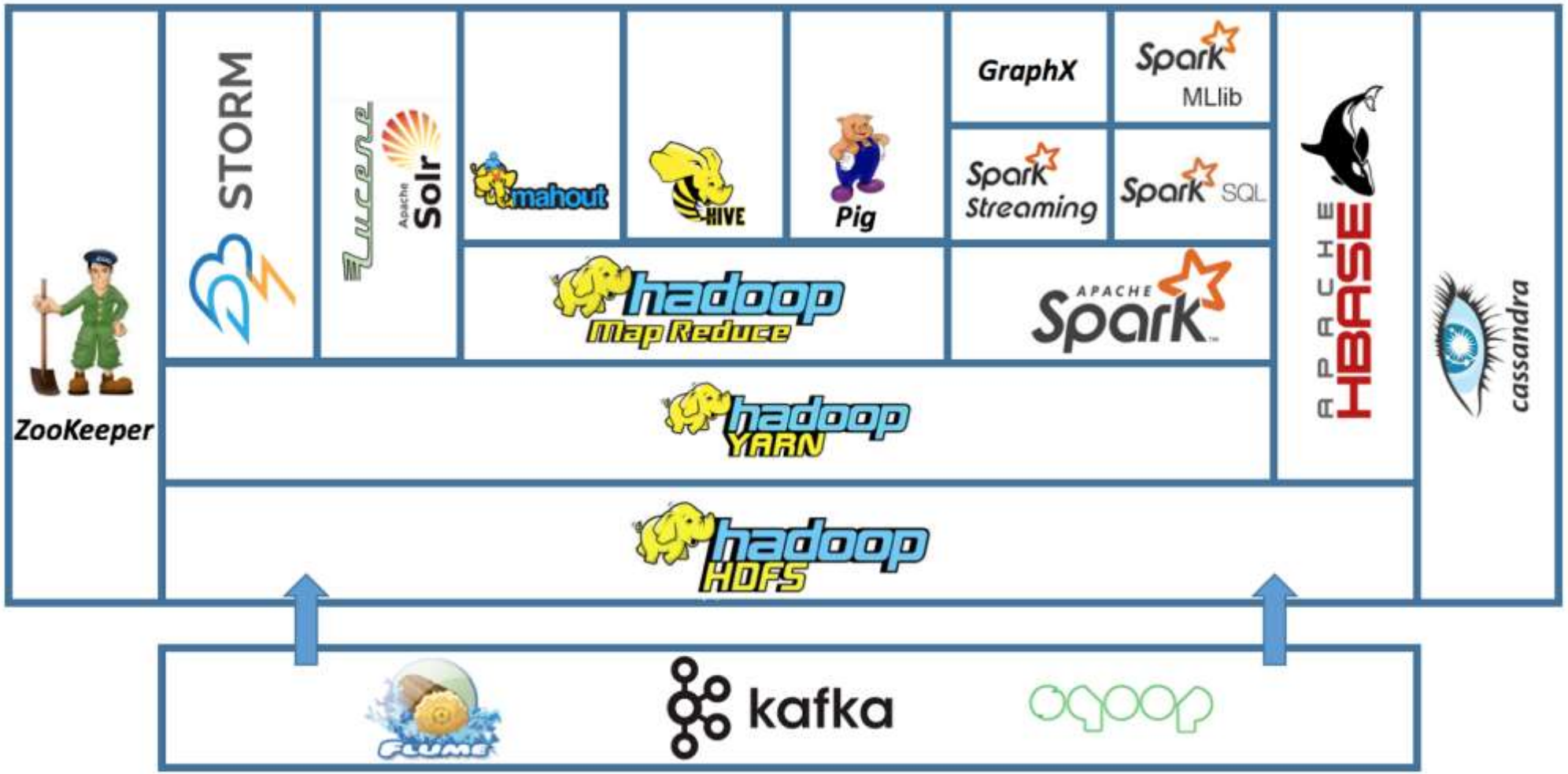
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Apache Hadoop e Big Data Stack

Apache Hadoop e Big Data Stack



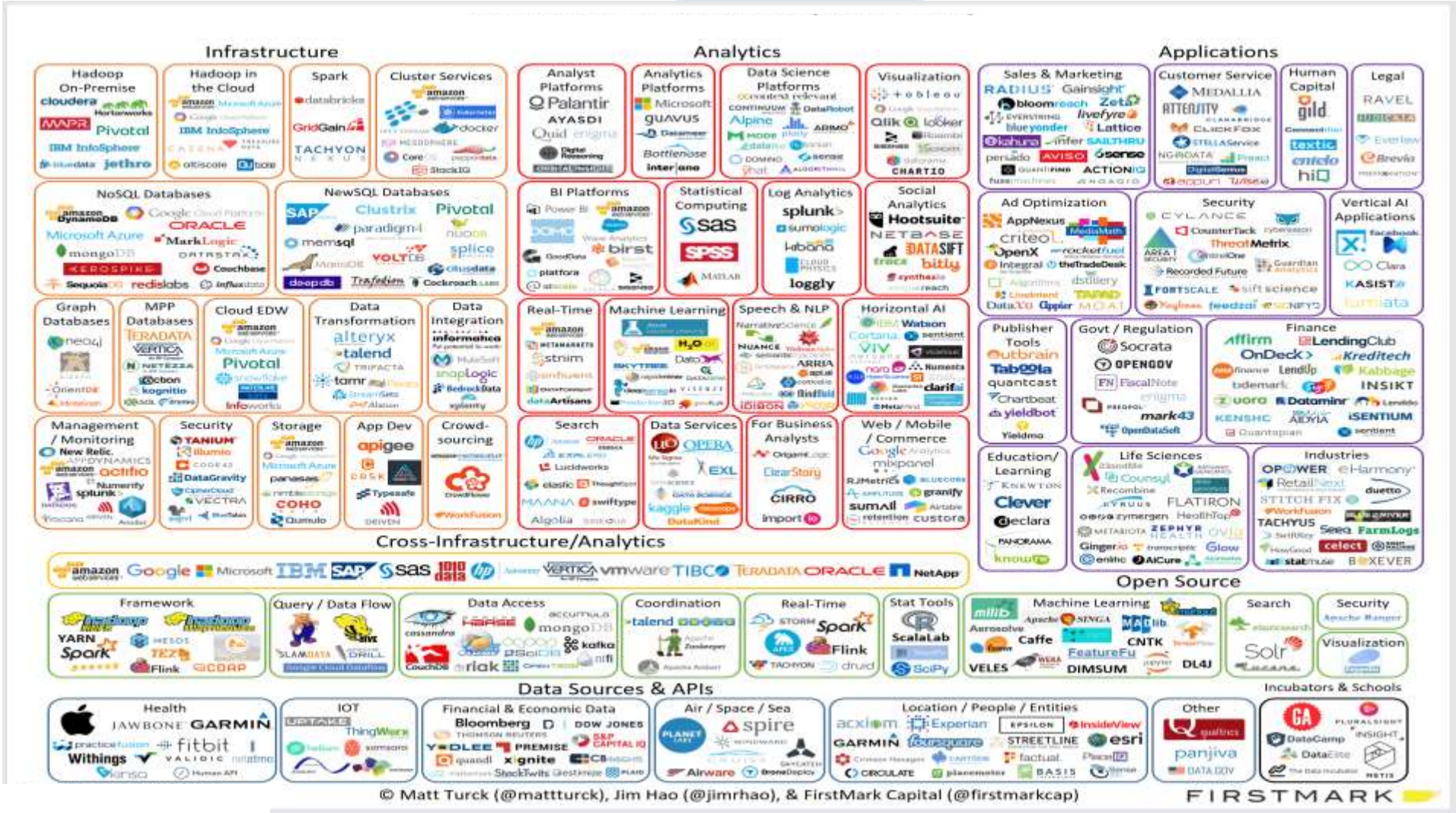
Apache Hadoop e Big Data Stack



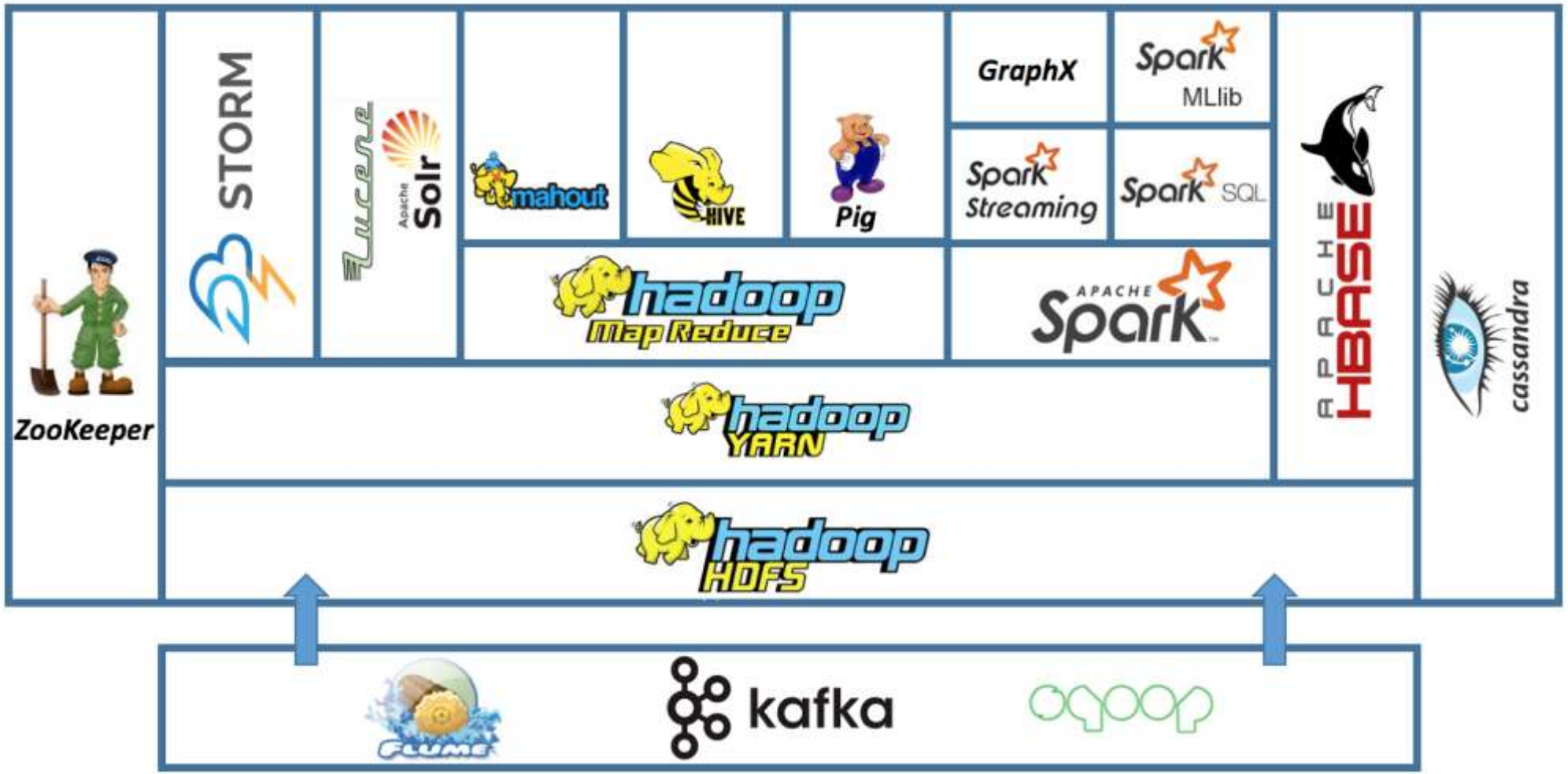
Apache Hadoop e Big Data Stack

The Big Data technology stack is evolving rapidly

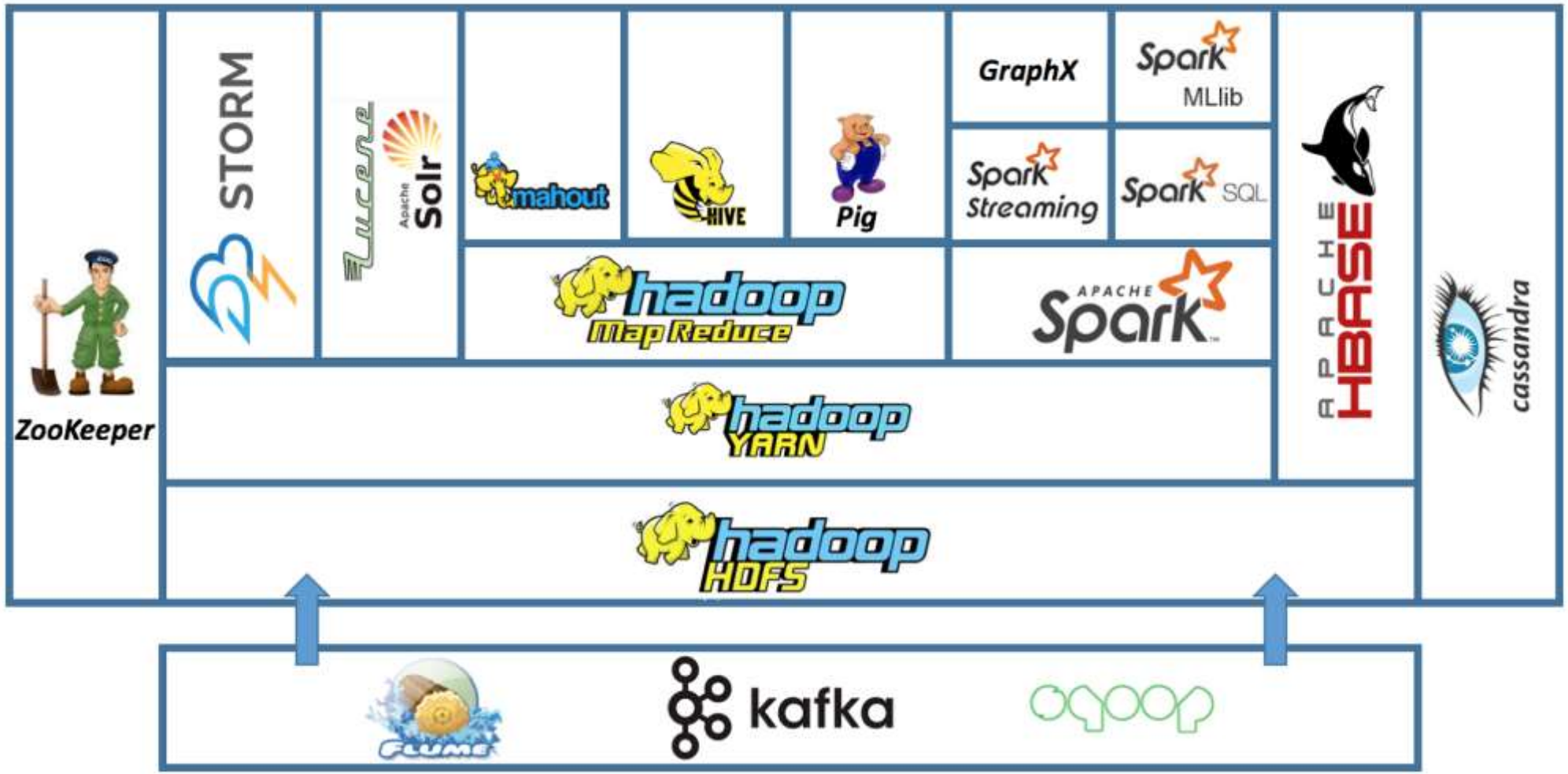
Apache Hadoop e Big Data Stack



Apache Hadoop e Big Data Stack



Apache Hadoop e Big Data Stack



Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Dados em Batch
x
Streaming de Dados

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

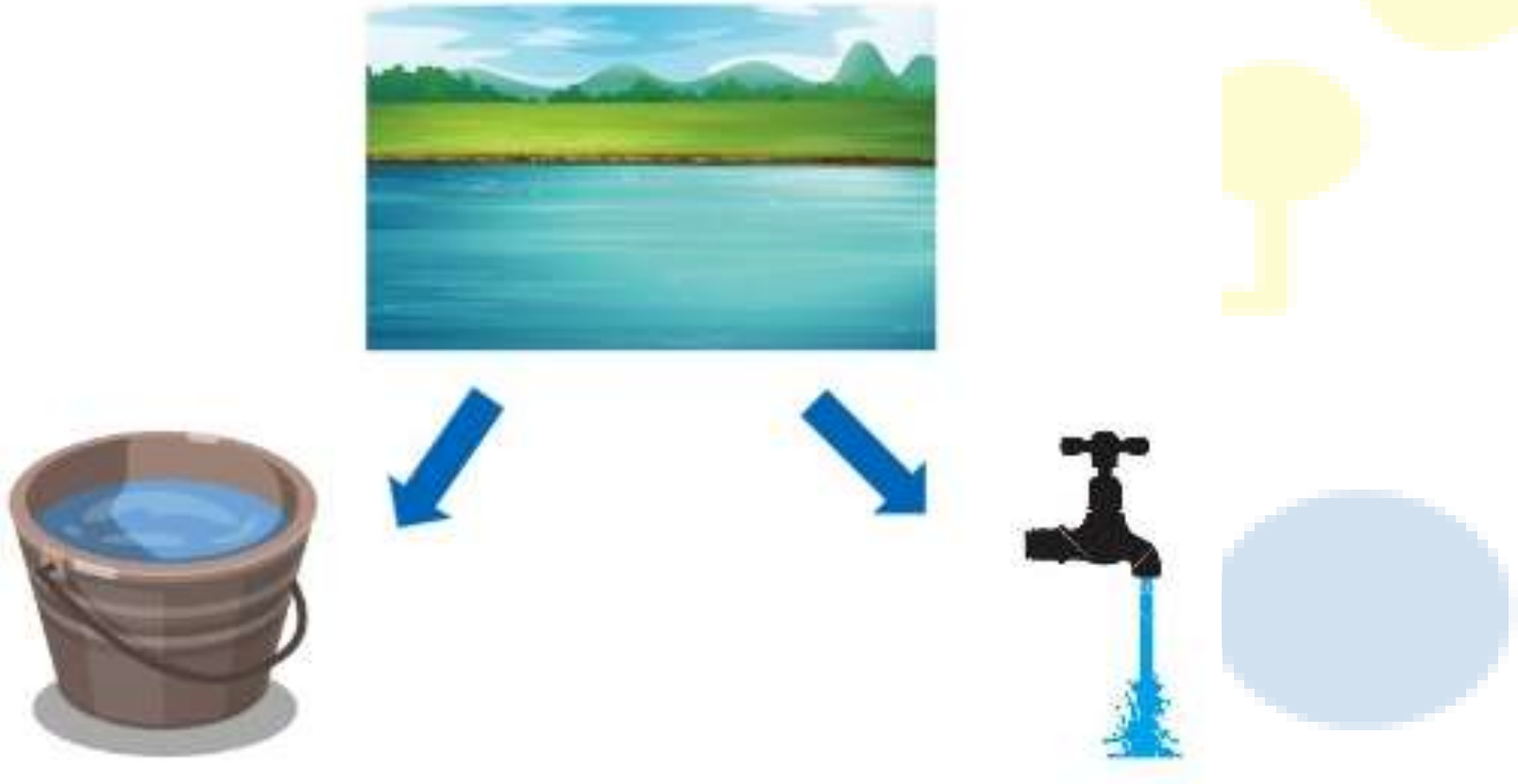
Dados em Batch x Streaming de Dados

Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Dados em Batch x Streaming de Dados

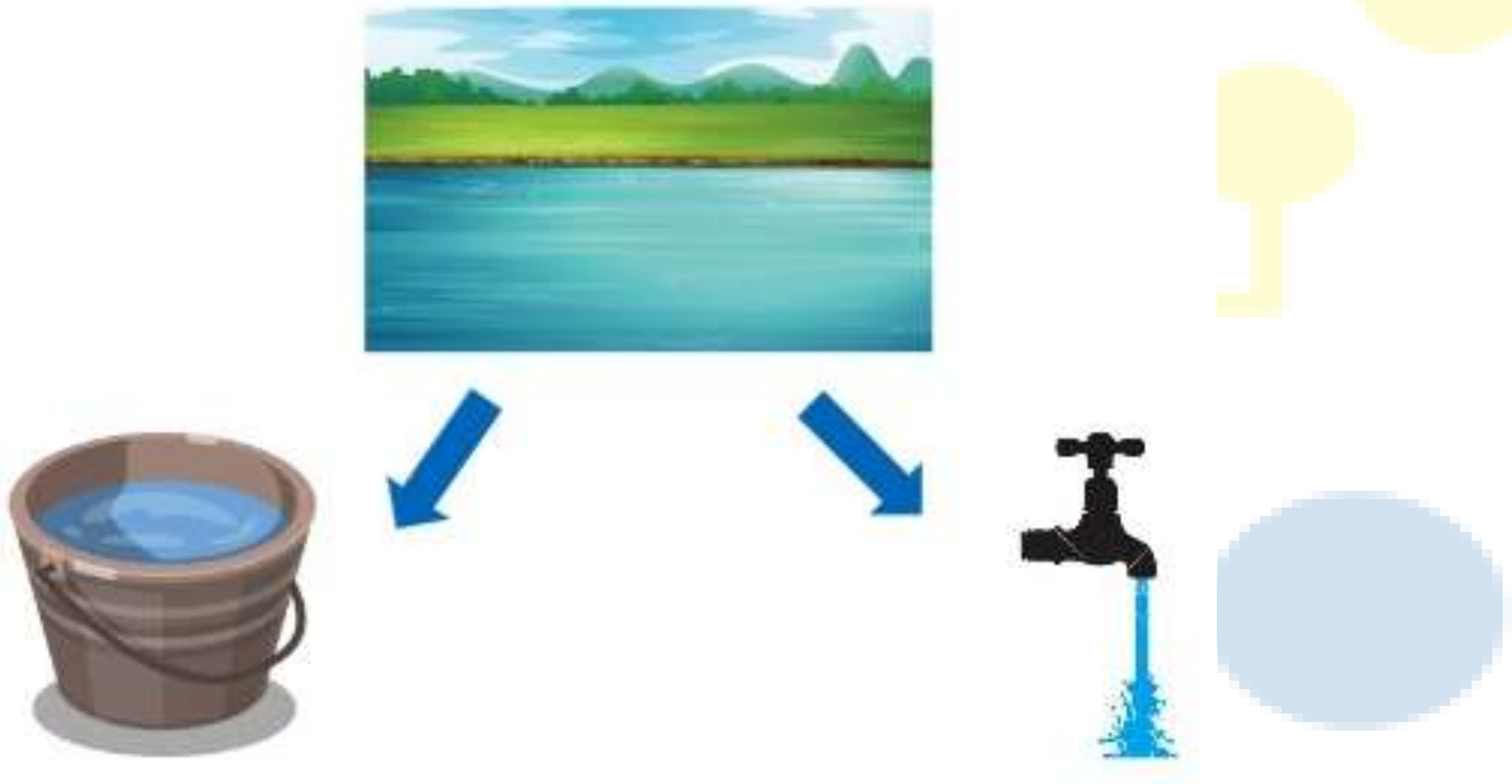
Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Dados em Batch x Streaming de Dados



Data Science Academy eng.davidborges@gmail.com 59532d8f5e4cdead748b456a

Dados em Batch x Streaming de Dados



Batch

Streaming

Muito Obrigado.

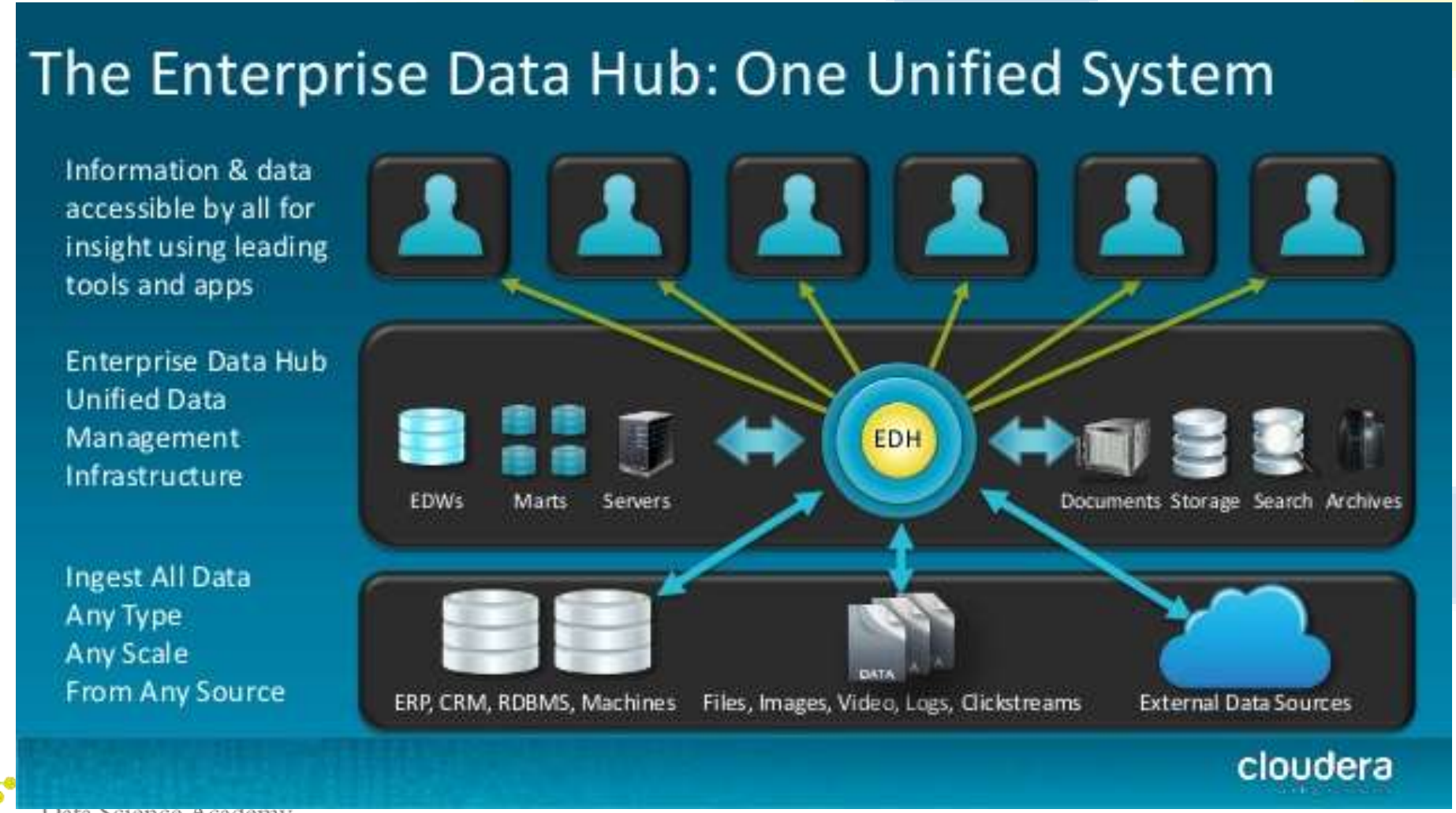
É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Arquitetura Corporativa Enterprise Data Hub

Arquitetura Corporativa - Enterprise Data Hub



Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.

DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO

Qual o Custo de Implementação de um Data Lake?

Qual o Custo de Implementação de um Data Lake On-premise?

Componente	Custo

Qual o Custo de Implementação de um Data Lake On-premise?

Componente	Custo
Cluster – 3 Masters (16 cores, 128 GB RAM, 2 x 2 TB)	~ USD $ 80.000
Cluster – 7 Nodes (16 cores, 256 GB RAM, 12 x 2 TB)	

Qual o Custo de Implementação de um Data Lake On-premise?

Componente	Custo
Cluster – 3 Masters (16 cores, 128 GB RAM, 2 x 2 TB)	~ USD $ 80.000
Cluster – 7 Nodes (16 cores, 256 GB RAM, 12 x 2 TB)	
Rede 10 Gbps, Softwares de Monitoramento, Custos de CPD, Storage, Switches, Suporte de Hardware	~ USD $ 25.000

Qual o Custo de Implementação de um Data Lake On-premise?

Componente	Custo
Cluster – 3 Masters (16 cores, 128 GB RAM, 2 x 2 TB)	~ USD $ 80.000
Cluster – 7 Nodes (16 cores, 256 GB RAM, 12 x 2 TB)	
Rede 10 Gbps, Softwares de Monitoramento, Custos de CPD, Storage, Switches, Suporte de Hardware	~ USD $ 25.000
Instalação, Configuração, Testes, Integração, Automatização	~ USD $ 35.000

Qual o Custo de Implementação de um Data Lake On-premise?

Componente	Custo
Cluster – 3 Masters (16 cores, 128 GB RAM, 2 x 2 TB)	~ USD $ 80.000
Cluster – 7 Nodes (16 cores, 256 GB RAM, 12 x 2 TB)	
Rede 10 Gbps, Softwares de Monitoramento, Custos de CPD, Storage, Switches, Suporte de Hardware	~ USD $ 25.000
Instalação, Configuração, Testes, Integração, Automatização	~ USD $ 35.000
Custo Anual Total	**~ USD $ 140.000**
Custo Mensal	**~ USD $ 12.000**

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo
Azure Data Lake Store (150 TB)	~ USD $ 5.500

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo
Azure Data Lake Store (150 TB)	~ USD $ 5.500
HDInsight Cluster (10 compute nodes)	~ USD $ 3.800

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo
Azure Data Lake Store (150 TB)	~ USD $ 5.500
HDInsight Cluster (10 compute nodes)	~ USD $ 3.800
Suporte Enterprise	~ USD $ 1.000

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo
Azure Data Lake Store (150 TB)	~ USD $ 5.500
HDInsight Cluster (10 compute nodes)	~ USD $ 3.800
Suporte Enterprise	~ USD $ 1.000
Instalação, Configuração, Testes, Integração, Automatização	~ USD $ 30.000

Qual o Custo de Implementação de um Data Lake na Nuvem?

Componente	Custo
Azure Data Lake Store (150 TB)	~ USD $ 5.500
HDInsight Cluster (10 compute nodes)	~ USD $ 3.800
Suporte Enterprise	~ USD $ 1.000
Instalação, Configuração, Testes, Integração, Automatização	~ USD $ 30.000
Custo no Primeiro Mês	**~ USD $ 40.300**
Custo Mensal	**~ USD $ 10.300**

Qual o Custo de Implementação de um Data Lake?

Custo Mensal On-premise
~ USD $ 12.000

Custo Mensal na Nuvem
~ USD $ 10.300

Muito Obrigado.

É um prazer ter você aqui.
Tenha uma excelente jornada de aprendizagem.