A aula interativa do Módulo 3 – Bootcamp Engenheiro(a) de Dados começará em breve!

Atenção:

1) Você entrará na aula com o microfone e o vídeo DESABILITADOS.

2) Apenas a nossa equipe poderá habilitar seu microfone e seu vídeo em momentos de interatividade, indicados pelo professor.

3) Utilize o recurso Q&A para dúvidas técnicas. Nossos tutores e monitores estarão prontos para te responder e as perguntas não se perderão no chat.

4) Para garantir a pontuação da aula, no momento em que o professor sinalizar, você deverá ir até o ambiente de aprendizagem e responder a enquete de presença. Não é necessário encerrar a reunião do Zoom, apenas minimize a janela.

Armazenamento de Dados

SEGUNDA AULA INTERATIVA

PROF. MARCILIO ANDRADE

Armazenamento de Dados

SEGUNDA AULA INTERATIVA

PROF. MARCILIO ANDRADE

Nesta aula

- Demonstração.
- Resolução de dúvidas.
- Revisão dos principais conceitos da segunda parte da disciplina (se o tempo permitir).
- Encerramento da disciplina.

Demonstração

https://media.istockphoto.com/vectors/demo-with-cursor-sign-icon-demonstration-symbol-vector-id610564868?k=6&m=610564868&s=170667a&w=0&h=rqPZtPJlt3x40KUV5LdX8foId_LsIUKbBq6o9Dv7dKw=

Resolução de Dúvidas

Encerramento da Disciplina

https://lh3.googleusercontent.com/proxy/IZDtB7I-wgJhA-hVQO_7VcKX0xcbm8R42QqrawRQf3MeXfIxY4nGDohSHoWZxJjxz65CSJYyzATzz8j3lYP5OPTXcg05uciZ_3iZ4KO23foCCWOD0nrpypj97qmbjxdCNA

Aplicação do Teorema CAP

Consistência

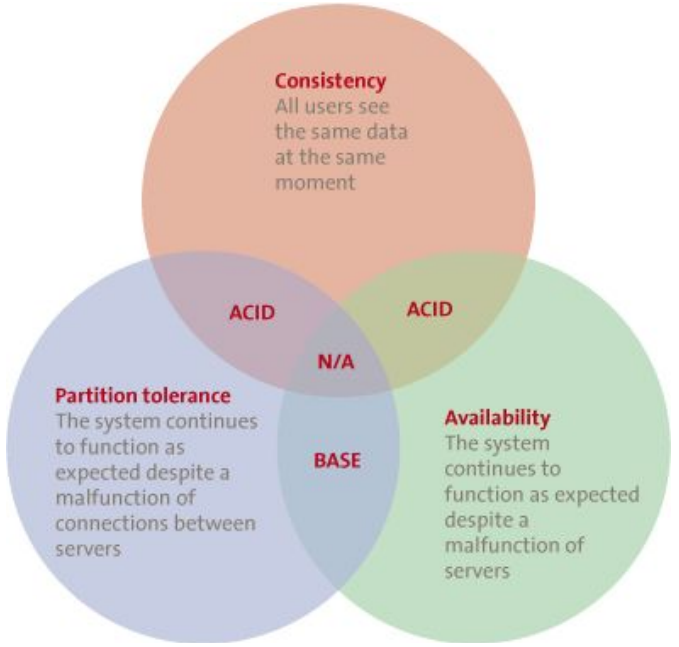
Todos os nós devem ver os mesmos ao mesmo tempo, o que significa que as réplicas devem ser atualizadas imediatamente.

Disponibilidade

Uma solicitação é sempre atendida pelo sistema. Nenhuma solicitação recebida é perdida.

Tolerância a Partição

O sistema continua a operar mesmo em caso de falha de um nó. O sistema falhará apenas se todos os nós falharem.



https://www.compact.nl/wordpress/wp-content/uploads/2016/02/C-2013-0-Keur-02.png

ACID x BASE

A	C	I	D
• *Atomicity*. • Atomicidade. • Unidade indivisível de trabalho. • Tudo feito ou nada feito.	• *Concistency*. • Consistência. • Restrições de integridades garantidas. • Dados coerentes.	• *Isolation*. • Isolamento. • Dado alterado por apenas uma transação por vez. • Espere, é a minha vez.	• *Durability*. • Durabilidade. • Garantia de persistência do dado. • Salvo com sucesso.

ACID x BASE

BA

- *Basically Available*.
- Basicamente disponível.
- Banco de dados disponível todo o tempo.
- Disponibilidade é fundamental.

S

- Soft-State.
- Estado suave (flexível).
- Não precisa ser consistente em tempo integral.
- Não estou consistente nesse momento, mas tudo bem.

E

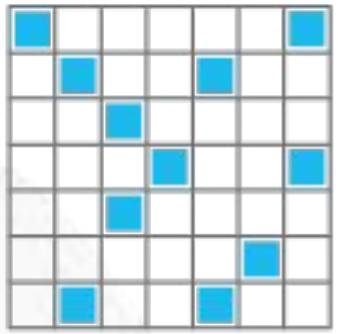
- *Eventually Consistent*.
- Eventualmente consistente.
- O banco passará a um estado consistente no momento devido.
- Estarei consistente em breve.

Pontos fortes e fracos

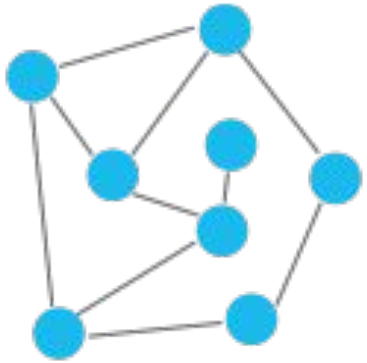
- Capacidade de tratar dados com volume, velocidade e variedade.
- Simplicidade do modelo de dados.
- Flexibilidade da estrutura de dados.
- Facilidade para escalar horizontalmente.
- Diferentes opções de armazenamento de dados.
- Investimento reduzido (código aberto).

- Ausência de ACID.
- Menor maturidade se comparado ao relacional.
- Menor segurança de autenticação e armazenamento.
- Menor quantidade de profissionais qualificados.
- Dificuldade de suporte profissional com garantia de SLA.

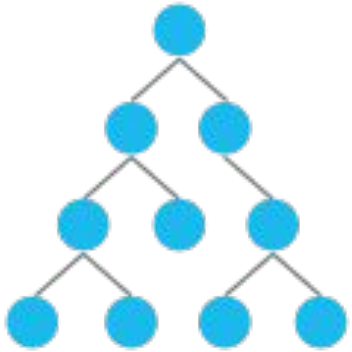
Modelos de NoSQL



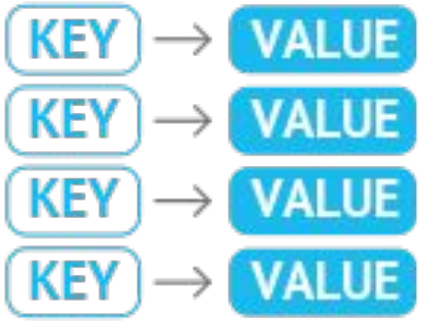
Column-Family



Graph



Document



Key-Value

https://devcom.com/wp-content/uploads/2020/06/nosql.png

https://www.amigoconstrutor.com.br/wp-content/uploads/2020/05/shutterstock_196063721.jpg

Object storage, file storage e block storage

Object storage	*File storage*	*Block storage*
• Armazenamento de objetos. • Metadados associados. • Identificador exclusivo. • Escalabilidade. • Dados para análise, backup ou arquivamento.	• Armazenamento de arquivos. • Armazenamento hierárquico (pastas). • Necessário saber o caminho físico para acessá-los. • Network Attached Storage (NAS). • Repositórios de conteúdo, ambientes de desenvolvimento, armazenamentos de mídia ou diretórios de usuários.	• Armazenamento de blocos. • Arquivo dividido em blocos singulares de dados. • Cada parte tem um endereço diferente. • Direct Attached Storage (DAS) e Storage Area Network (SAN). • Bancos de dados ou sistemas ERP que exigem um armazenamento dedicado e de baixa latência.

Comparativo entre fornecedores

	AWS	Azure	Google Cloud
Archival storage	S3 Glacier, S3 Glacier Deep Archive	Archive Storage	Archive Storage
Backup	AWS Backup	Azure Backup	N/A
Block storage	Amazon Block Store (EBS)	Disk Storage	Persistent Disk, Local SSD
File storage	Amazon Elastic File Service (EFS), Amazon FSx for Windows File Server, Amazon FSx for Lustre	File Storage, Azure NetApp Files	Filestore
Object storage	Amazon S3	Blob storage	Cloud Storage, Cloud Storage for Firebase

https://searchcloudcomputing.techtarget.com/feature/A-cloud-services-cheat-sheet-for-AWS-Azure-and-Google-Cloud

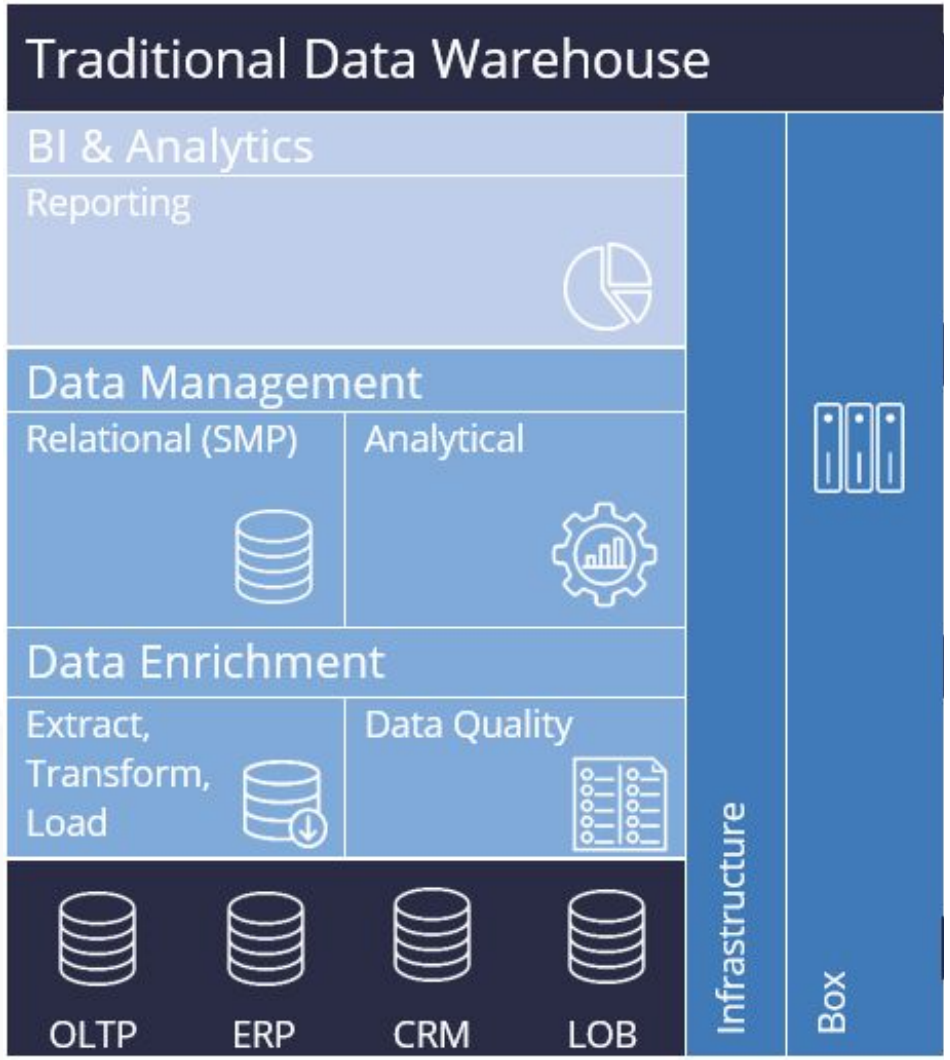
AWS S3 versus HDFS

- Armazenamento em nuvem fornecem elasticidade, com melhor disponibilidade e durabilidade, maior desempenho e menor custo que clusters HDFS tradicionais.
- O HDFS transformou em *commodity* o armazenamento de big data, tornando mais barato armazenar e distribuir uma grande quantidade de dados.
- No entanto, em uma arquitetura nativa da nuvem, o benefício do HDFS é mínimo e não compensa a complexidade operacional.

	S3	HDFS	S3 vs HDFS
Elasticidade	sim	Não	S3 é mais elástico
Custo / TB / mês	$ 23	$ 206	10X
Disponibilidade	99,99%	99,9% (estimado)	10X
Durabilidade	99,999999999%	99,9999% (estimado)	10X +
Gravações transacionais	Sim com DBIO	sim	Comparável

https://databricks.com/blog/2017/05/31/top-5-reasons-for-choosing-s3-over-hdfs.html

DW: Tradicional x Moderno

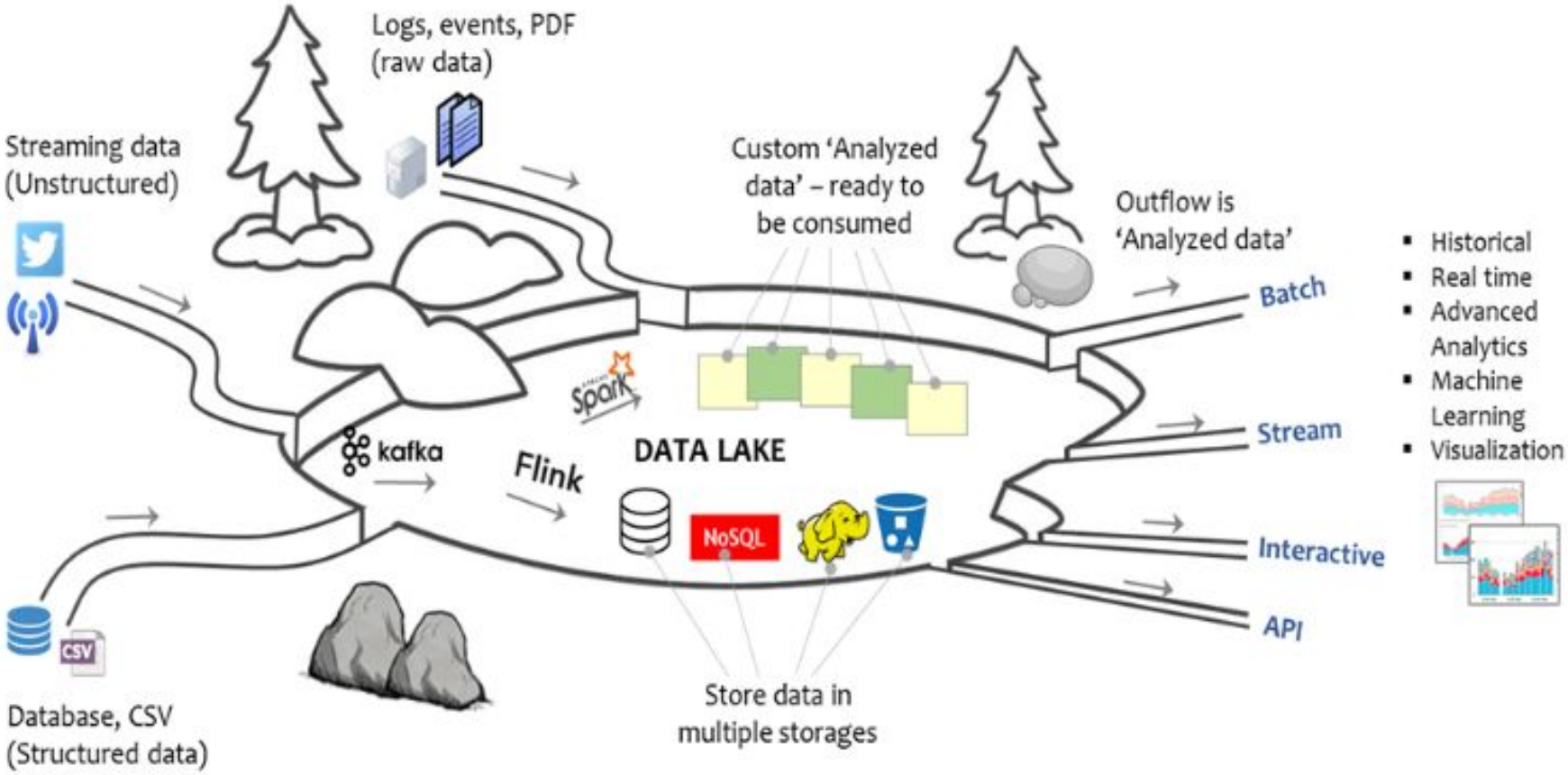


https://www.predicagroup.com/app/uploads/2020/08/Modern-Data-Warehouse-model.png

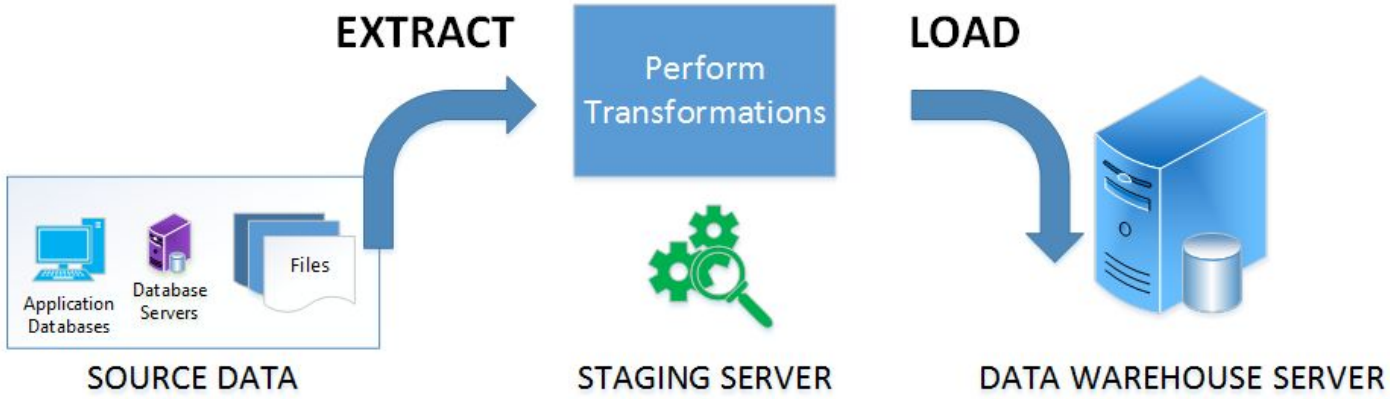
DW: Tradicional x Moderno

https://www.predicagroup.com/app/uploads/2020/08/Modern-Data-Warehouse-model.png

Estrutura de um *Data Lake*

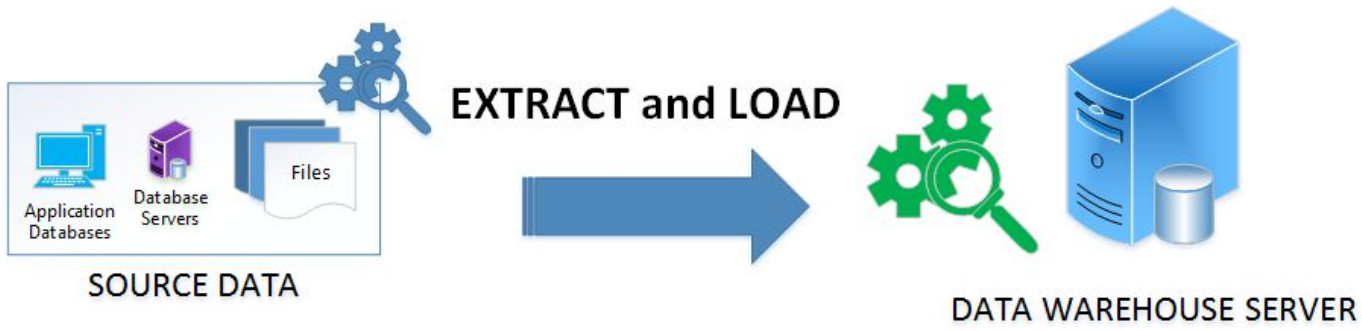


ETL x ELT

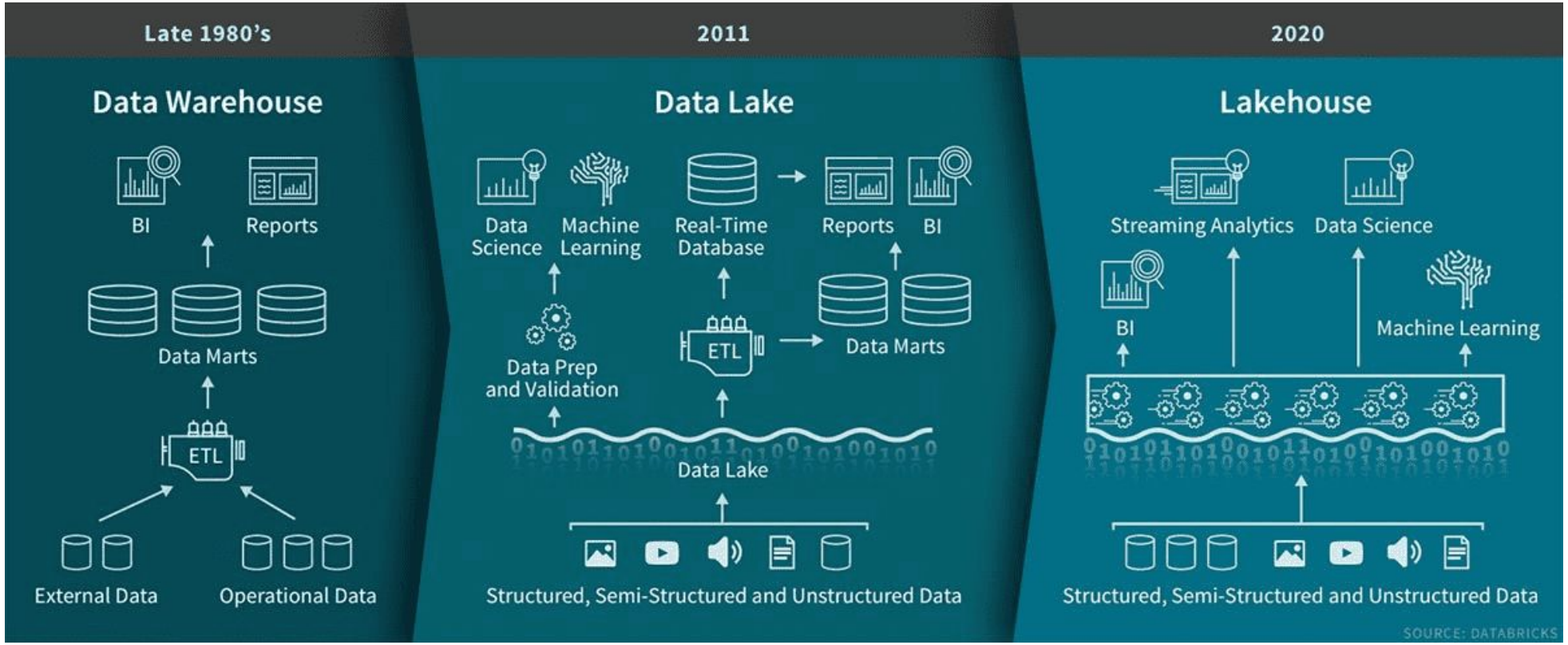


ETL (Extract – Transform – Load)

ELT (Extract – Load - Transform)



Data Lakehouse



https://databricks.com/wp-content/uploads/2020/01/data-lakehouse.png

Abordagens de Virtualização de Dados

APPROACH	DESCRIPTION	PROS	CONS	EXAMPLE VENDORS
Core Data Virtualization Platforms	A stand alone virtual data (semantic) layer that abstracts the physical data platform and location and allows queries across disparate data platforms.	Platform independent True semantic layer Caching for performance	Separate service to manage Data needs to be modeled	AtScale Denodo Dremio TIBCO DV
SQL-On-Anything	A query federation engine that allows a single SQL query to combine data from more than one data platform.	Good for file-based access Scale out with clustering	Users need to understand remote schemas Unpredictable performance	Amazon Athena Apache Drill Presto
Remote Data Source Bridges	A database extension that is embedded in a RDBMS that allows SQL access to remote databases using "external" tables.	Good when one primary ware-house is used No separate service to manage	Users need to understand remote schemas Unpredictable performance	Amazon Redshift Spectrum IBM Db2 Big SQL Microsoft SQL Server Polybase Oracle Big Data SQL Teradata QueryGrid
Autonomous Data Warehouses	A platform that automates the modeling, integration, and con-nectivity of source data which is then loaded into a target data platform.	Data stays in one place Performance is easier to predict	Requires data movement (data copies) Data latency or staleness	Data Virtuality Infoworks.io Incorta

https://www.atscale.com/wp-content/uploads/2020/01/Complete_Buyers_January21.pdf

Conclusão

- ✔ Realizada demonstração de mais uma ferramenta capaz de realizar a virtualização de dados.
- ✔ Sanadas possíveis dúvidas.
- ✔ De acordo com a disponibilidade de tempo, apresentada revisão dos principais conceitos da primeira parte da disciplina.
- ✔ Realizado debate em grupo.
- ✔ Encerramento da disciplina.