Bancos de Dados Tradicionais
Data Warehouses e Data Lakes

Data Warehouse

- Os **bancos de dados tradicionais**, como aqueles relacionais vistos anteriormente, são chamados de bancos de dados **transacionais**, pois:
  - São projetados para o processamento rotineiro de transações, suportando inclusões constantes de novos registros, exclusões e atualizações corriqueiras dos dados.
  - Equilibram a exigência de velocidade de acesso aos dados com a necessidade de assegurar a integridade deles.

Data Warehouse

- **Definição:**
- W. H. Inmon caracterizou um Data Warehouse como:

"Uma coleção de dados orientada a assunto, integrada, não volátil, variável no tempo para o suporte às decisões da gerência"

- **Características:**
  - Mudam com menos frequência (**atualização periódica**) e não podem ser consideradas de tempo real.
  - **Menos detalhada** e atualizada de acordo com uma escolha cuidadosa de politica de atualização

Data Warehouse

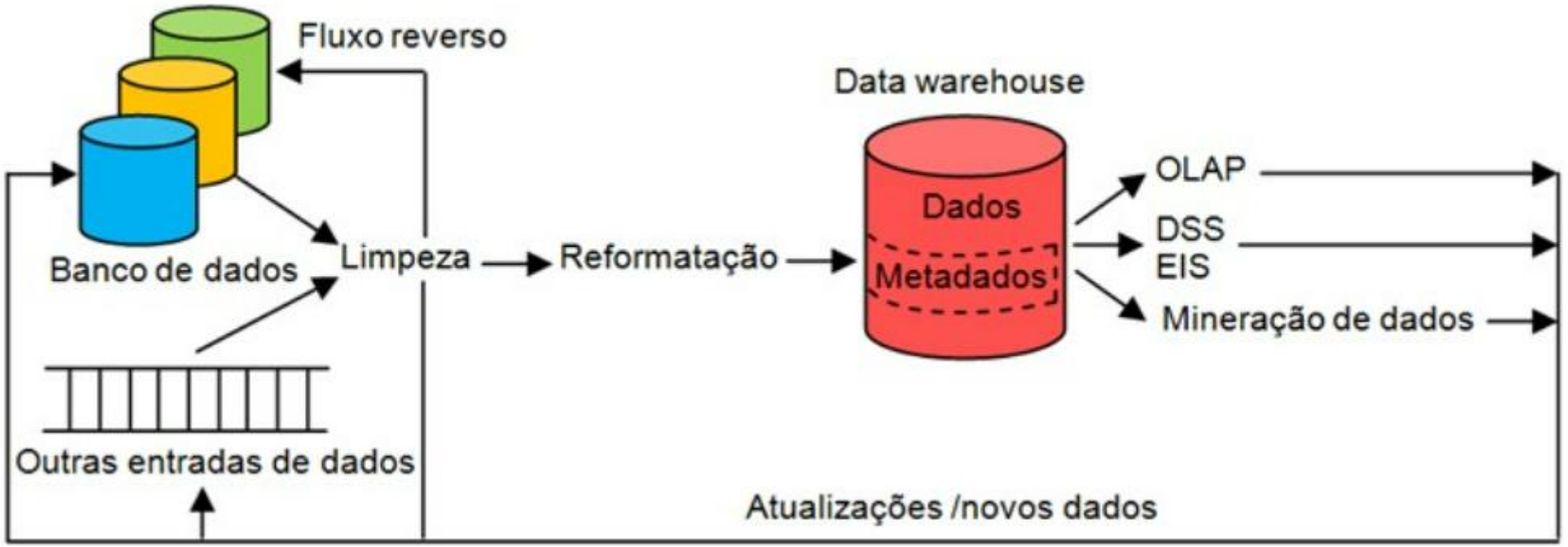
- Os ***data warehouses*** são otimizados para a recuperação de dados, **não** para o **processamento rotineiro de transações**.
- Proporcionam acesso aos dados para **análise complexa**, **descoberta de conhecimento** (mineração de dados) e **tomada de decisão**.
- Geralmente contêm **quantidades muito grandes de dados**, oriundos de **múltiplas fontes**.
- Diferentes fontes de dados tipicamente armazenam os dados usando representações e formatos diferentes, que necessitam ser padronizados para compor o warehouse.
- Suporte à decisão pode necessitar de **dados históricos**, os quais geralmente **não são mantidos pelos bancos transacionais**, mas são incorporados nos *data warehouses*.

Data Warehouse

- Os **bancos de dados tradicionais** dão apoio ao processamento *on-line* de transações (*online transaction processing – OLTP*), com **inclusões, atualizações e exclusões** frequentes dos dados.
- Os ***data warehouses*** dão suporte:
  - À **sistemas de apoio a decisão** (DSS ou EIS) com dados de nível mais alto e decisões complexas e importantes.
  - Ao **processamento analítico on-line** (OLAP – *online analytical processing*)
  - À **mineração de dados**

Data Warehouse

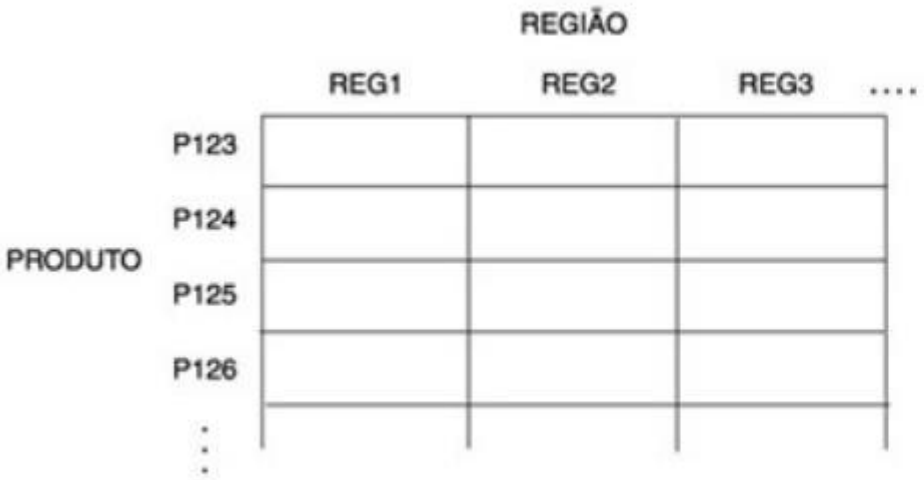
- **Estrutura Conceitual de um Data Warehouse**



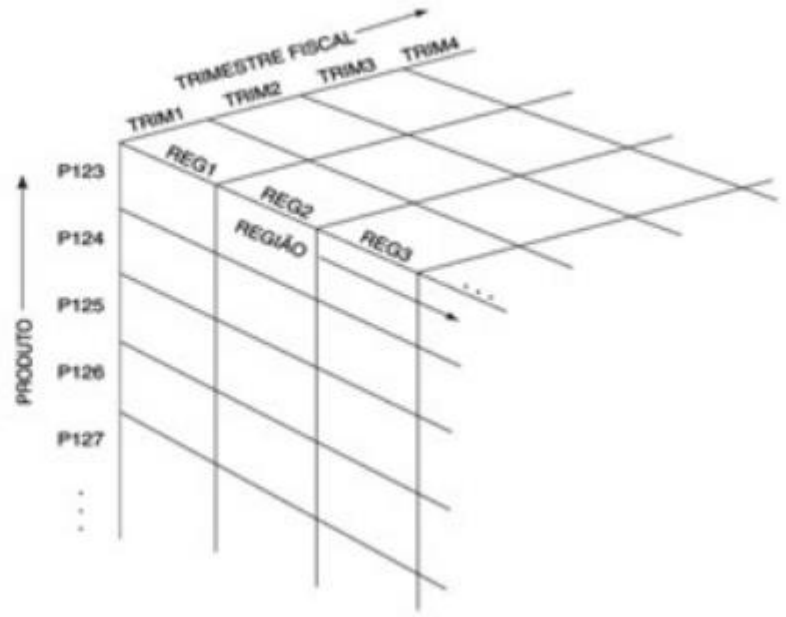
Modelagem Mutidimensional

- **Modelos Multidimensionais:** transformam os relacionamentos dos dados em matrizes multidimensionais, denominados **cubos de dados**

Modelo de Matriz Bidimensional

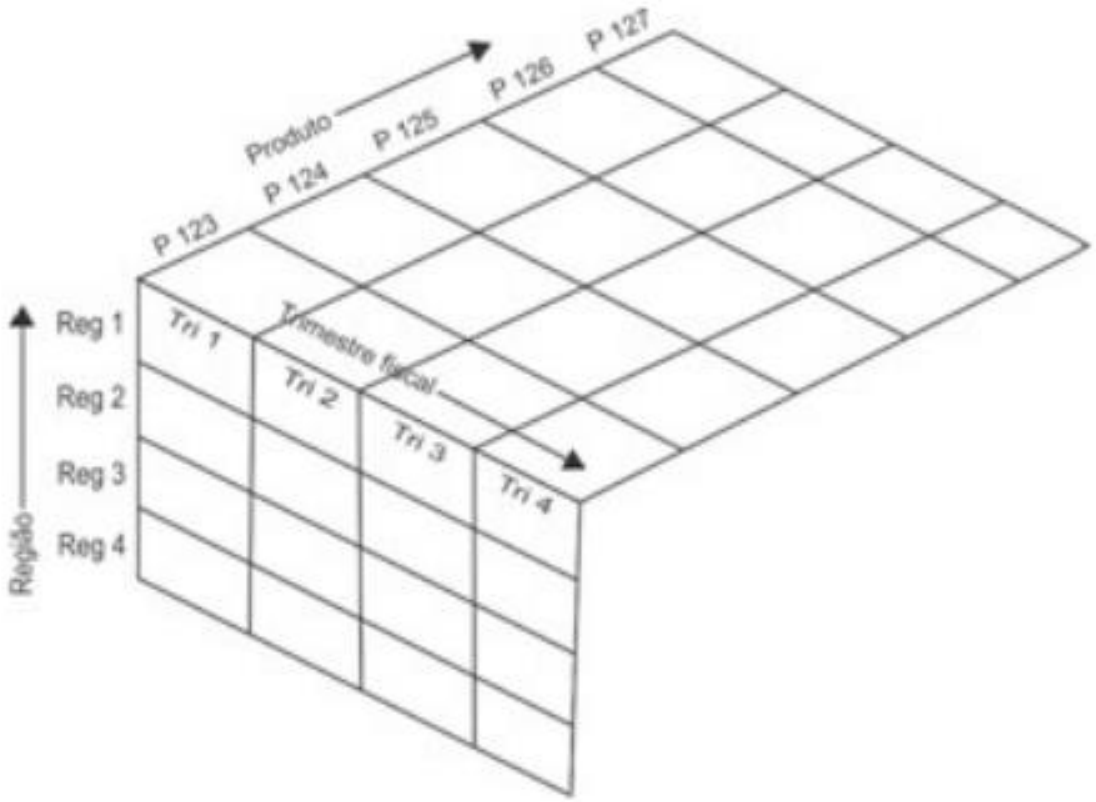


Modelo de Cubo de Dados Tridimensional



Modelagem Mutidimensional

- Mudar a hierarquia (orientação) unidimensional para outra é um processo fácil em cubos de dados, técnica denominada **giro**.

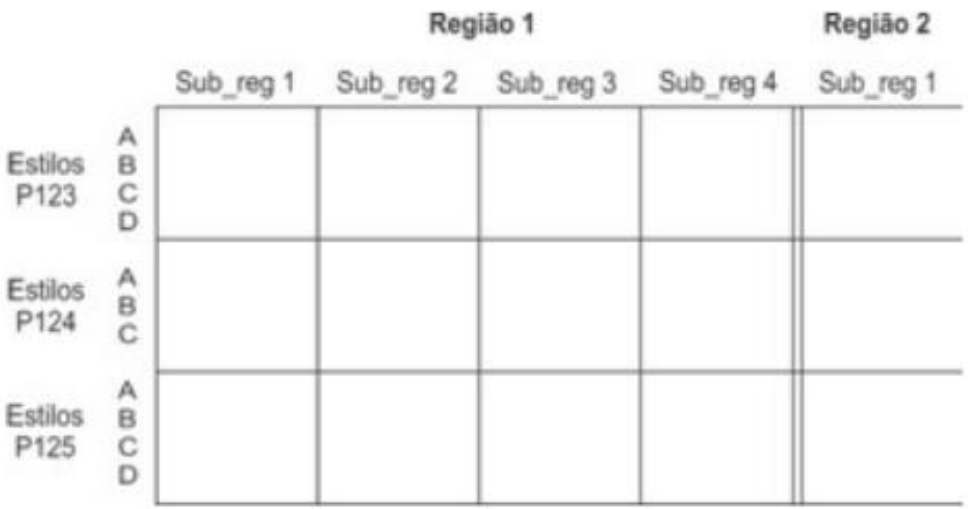


Modelagem Mutidimensional

- Modelos Multidimensionais atendem prontamente a visões hierárquicas no que é conhecido como exibição *roll-up* ou *drill-down*.

Operação *roll-up*

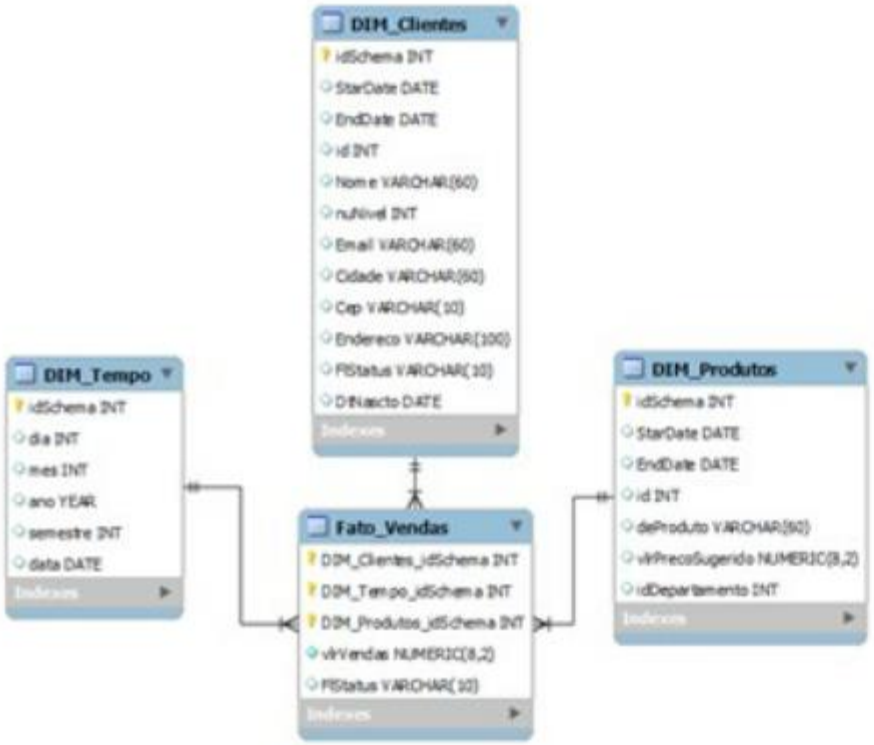
Operação *drill-down*



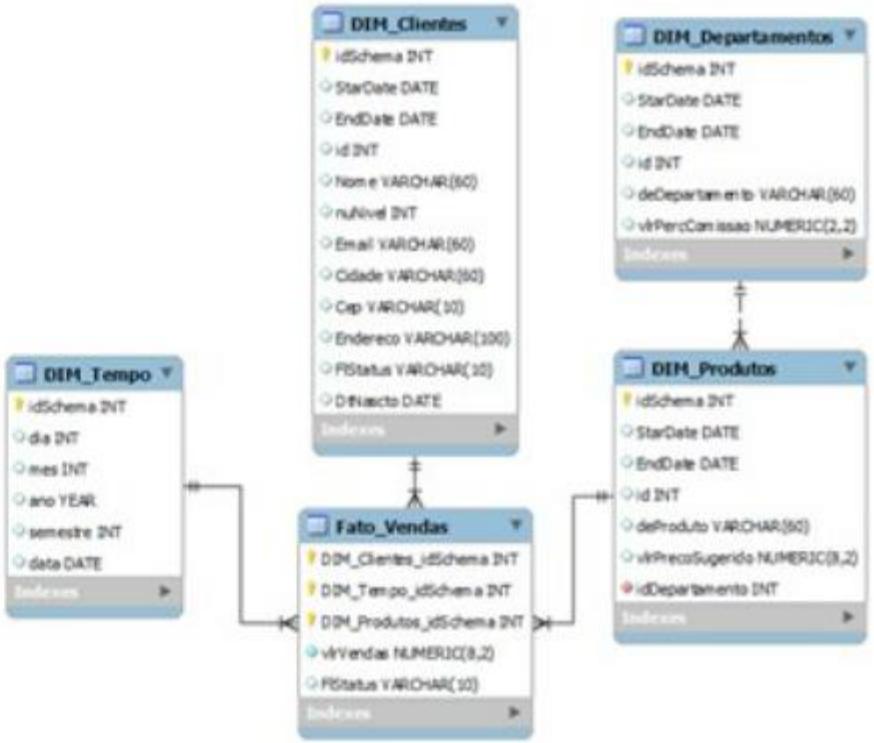
Modelagem Mutidimensional

- O **Star Schema** e o modelo **SnowFlake** são dois esquemas comuns para o design de *data warehouses*.

Star Schema



Snow Flake



Modelagem Mutidimensional

- A tabela de **fato** se relaciona com diversas tabelas **dimensão** no Star Schema, através de múltiplas junções por meio de uma chave primária composta.
- As tabelas **dimensão**, por sua vez, geralmente compostas de chaves primárias simples.
- A desnormalização das tabelas dimensão no Star Schema pode gerar a presença de dados altamente redundantes.
- A redundância no Star Schema é fundamental para melhoria no desempenho das consultas, visto que menos junções são fundamentais para recuperação dos dados.

Modelagem Mutidimensional

- O **SnowFake** é uma variação do esquema modelo Star Schema em que as tabelas dimensões de um esquema estrela são organizadas em uma hierarquia ao normalizá-las.
- Os benefícios da normalização, como a eliminação de redundâncias, geralmente comprometem o desempenho das consultas no *data warehouse*.

Modelagem Mutidimensional

- Uma **tabela de fato** pode ser imagina como tendo tuplas, uma para cada fato registrado.
- As tabelas fato são o ponto focal de um modelo dimensional, em que os dados de medição numérica são armazenados.
- Uma **constelação de fatos** é um conjunto de tabelas de fatos que compartilham algumas tabelas de dimensão.

Constelação de fatos

Modelagem Mutidimensional

- Uma **tabela dimensão** consiste em tuplas de atributos da dimensão.
- As tabelas dimensão sempre se **relacionam** com as **tabelas fato** e contêm as características de um evento.
- Como **exemplo** de tabelas dimensão de uma empresa do varejo, podemos mencionar *Tempo*, *Produto* ou até mesmo *Clientes*.

Modelagem Mutidimensional

- **Slowly Changing Dimension**: são os grupos de dados que se alteram em ciclos de tempo maiores e de maneira irregular.
- Por exemplo: um cliente muda de cidade e passa a realizar compras com outro representante da empresa em sua nova região.
- *Slowly Changing Dimension* são diferenciadas em dois tipos ou níveis principais: o **tipo 2** e **tipo 6**.

Modelagem Mutidimensional

- **Slowly Changing Dimension (Tipo 2)**: envolve o registro de informações históricas, guardando uma linha para cada versão dos registros, fazendo uso das chaves substitutivas (*surrogate keys*).

Código	Fornecedor	Nome	Cidade	Ativo
1	1236	CompreTudo Ferragens	Florianópolis	0
2	1236	CompreTudo Ferragens	São José	1

Código	Fornecedor	Nome	Cidade	Data Inicial	Data Final
1	1236	CompreTudo Ferragens	Florianópolis	1/1/2008	1/10/2010
2	1236	CompreTudo Ferragens	São José	2/10/2010	Null

Modelagem Mutidimensional

- **Slowly Changing Dimension (Tipo 6)**: utiliza as duas metodologias do tipo 2 combinadas – colunas de data inicial e final e um campo booleano que determina se o registro está ativo ou não.

Código	Fornecedor	Nome	Cidade	Ativo	Data Inicial	Data Final
1	1236	Industria 01	Florianópolis	0	1/1/2008	1/10/2010
2	1236	Industria 01	São José	1	2/10/2010	Null

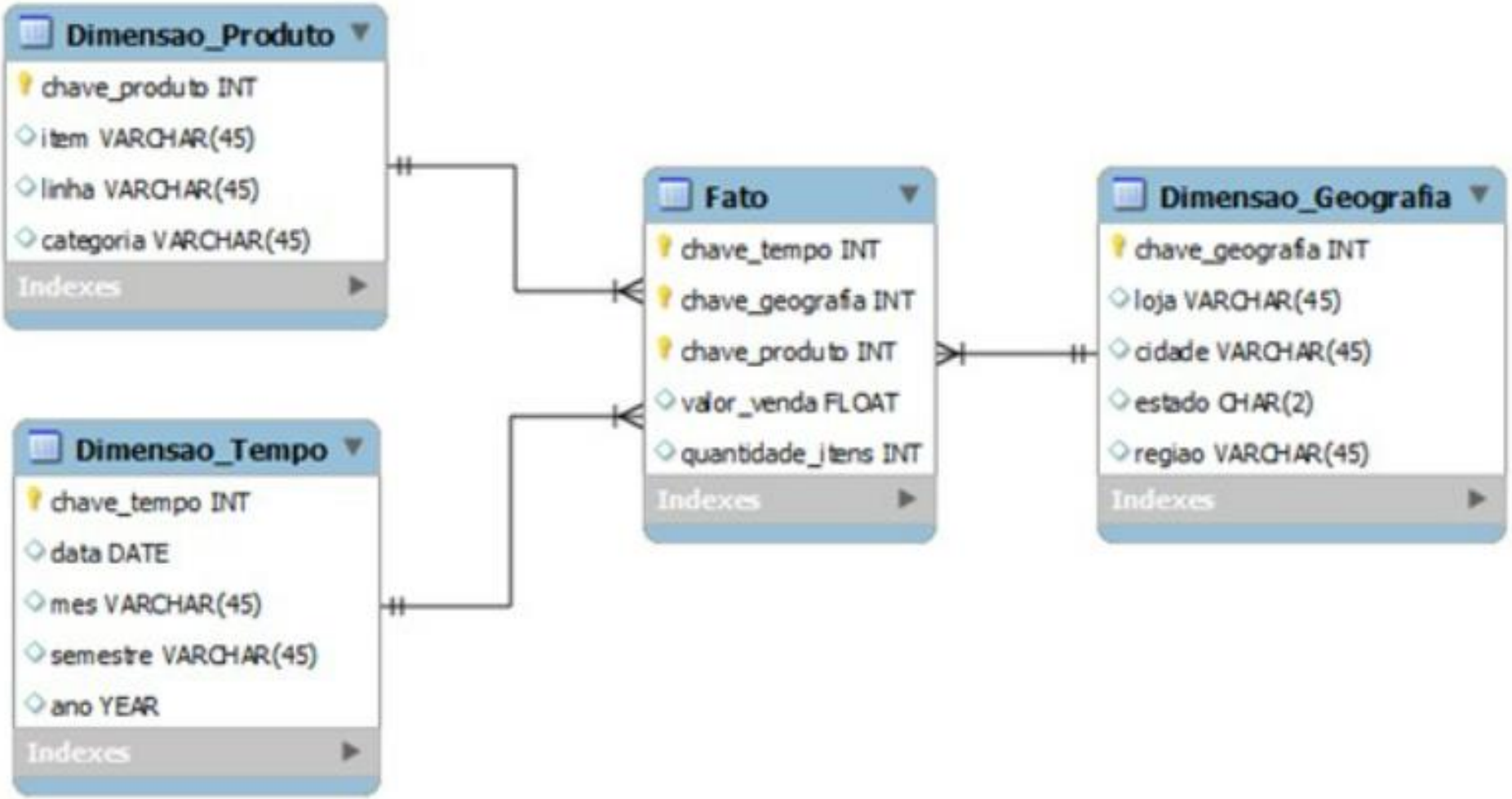
Modelagem Mutidimensional

- **Exercício 1:** Dado uma planilha de vendas de uma empresa do varejo, crie um modelo *Star Schema*

	A	B	C	D	E	F	G
1	Tempo (Dia)	Produto (Item)	Geografia (Loja)	Valor da venda (R$)	Quantidade de itens	Valor de venda (R$)	...
2							
3	05/01/04	Lápis n? 2 – Faver Carel	Loja 04	78	65	1,2	...
4	05/01/04	Lápis n? 2 – Faver Carel	Loja 06	150	125	1,2	...
5	05/01/04	Caneta Clic azul - fina	Loja 04	117,6	84	1,4	...
6	05/01/04	Caneta Clic vermelha - fina	Loja 04	39,2	28	1,4	...
7	...	...	...	...	...	...	...
8	23/03/04	Caneta Clic azul - fina	Loja 06	123	82	1,5	...
9	23/03/04	Bloco recibo Jordel	Loja 12	132,5	53	2,5	...
10	...	...	...	...	...	...	...

Modelagem Mutidimensional

- **Resposta do Exercício 1:**



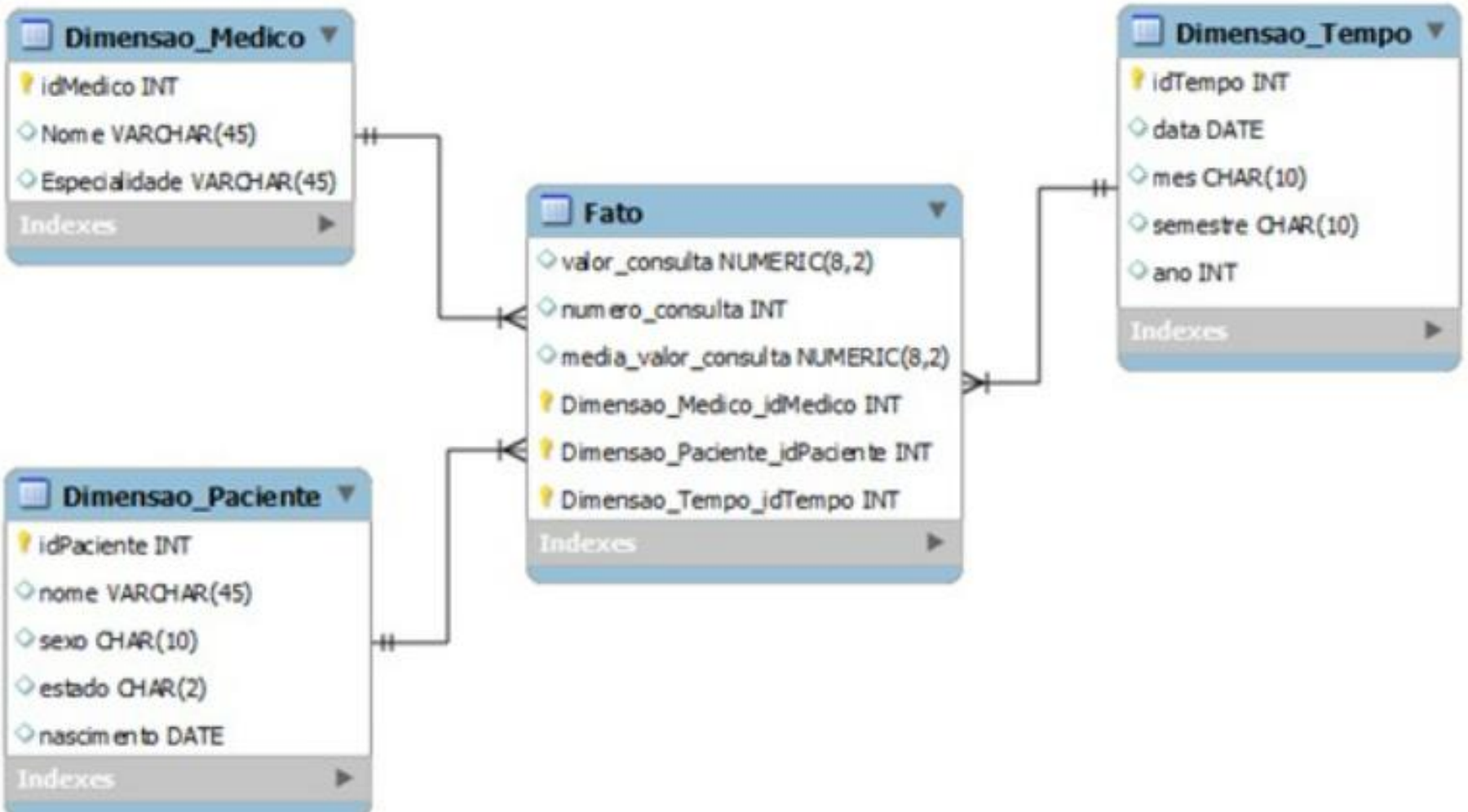
Modelagem Mutidimensional

- **Exercício 2:** Dado uma planilha de consultas diárias de uma clínica médica, crie um modelo *Star Schema*

	A	B	C	D	E	F	G	H	I
1	Data	Valor_consulta	Médico	Especialidade	Paciente	Sexo	Data Nasc	Estado	...
2	05/01/11	125.50	Carla Beker	Ginecologia	Carla Rocha	Feminino	03/06/82	BA	...
3	23/03/11	111.67	Pedro Zanuncio	Dermatologia	Bruna Oliveira	Feminino	25/08/85	PR	...
4	03/07/11	124.47	Domingos	Pediatria	Caetano Queiroz da Silva	Masculino	12/10/10	MS	...
5	17/05/12	62.18	Ricardo Guirelli	Clinico Geral	Paulo Gomes	Masculino	10/04/80	SP	...
6	...	...	...	...	...	...	...	...	...
7	18/09/12	268.00	Andyane Tetila	Infectologia	Renato de Melo	Masculino	15/11/72	MS	...
8	08/10/13	141.18	Pedro Zanuncio	Dermatologia	Bruna Oliveira	Feminino	25/08/85	PR	...
9	...	...	...	...	...	...	...	...	...

Modelagem Mutidimensional

- **Resposta do Exercício 2:**



Vantagens de um Data Warehousing

- **Alta performance** para **consultas complexas**.
- Um **alto processamento de uma consulta** complexa no ***data warehouse*** **não afeta o desempenho dos bancos** de dados **operacionais**.
- **Pode operar** quando os **bancos de dados fontes** estão **indisponíveis**.
- Contém **informação extras e integradas**, como **informação histórica**.

Bancos Transacionais x Data Warehouses

Bancos de dados transacionais (para OLTP)	*Data Warehouses* (para OLAP)
Poucos registros acessados por vez	Grandes volumes acessados por vez (milhões)
Acesso para consulta e atualização	Basicamente consultas
Tamanho banco de dados: 100 MB – 100 GB	100 GB – poucos terabytes
Milhares de usuários	Centenas de usuários
Sem redundância de dados	Dados redundantes podem estar presentes

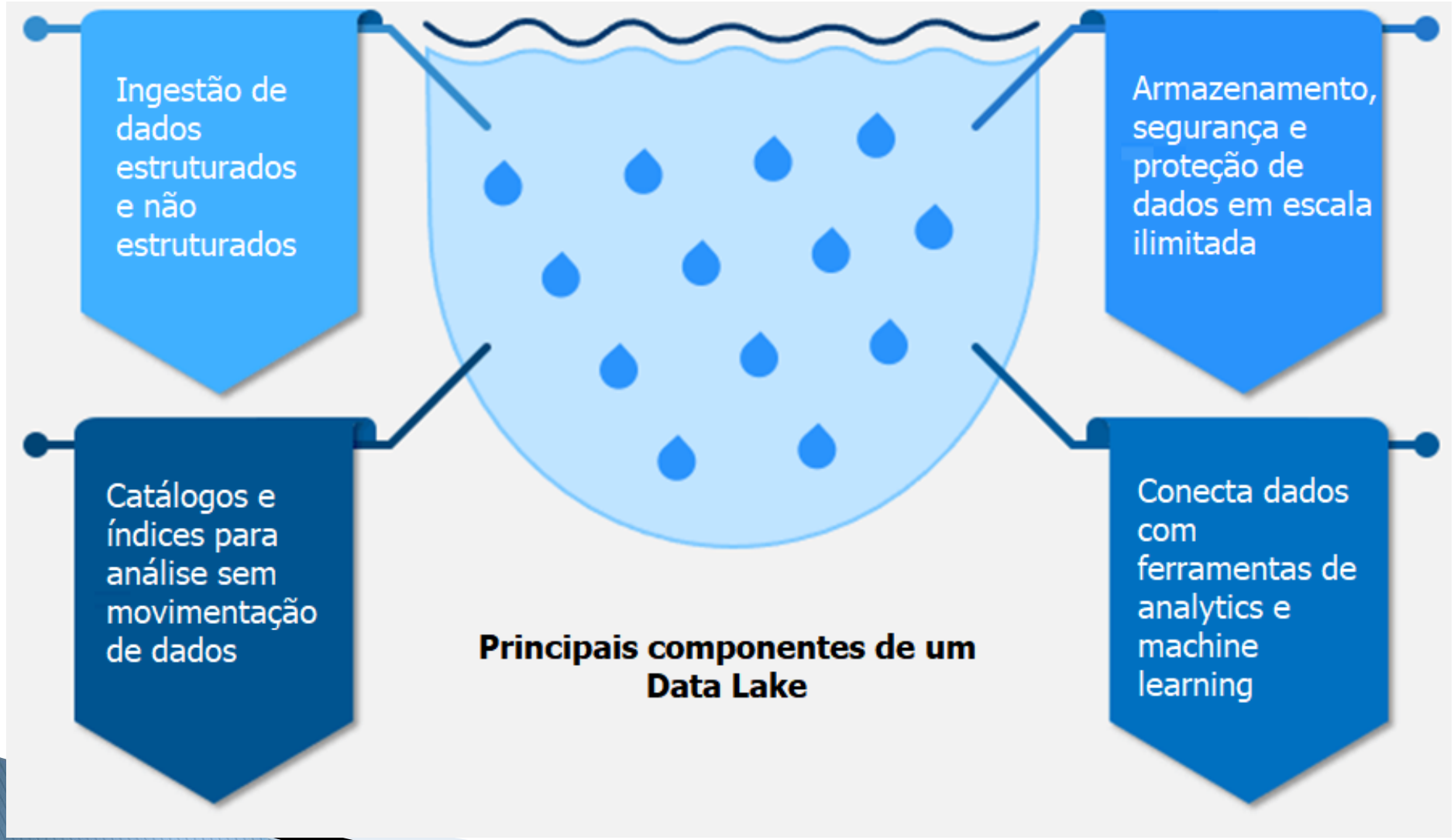
Data Mart

- Empresas podem montar *data warehouses* de âmbito empresarial para atender à toda organização.
- Ou podem criar **repositórios de dados menores**, **descentralizados**, denominados ***data marts*.**
- Assim, um ***data mart*** pode ser considerado uma parte, ou **subconjunto lógico** do *data warehouse* completo.
- Em geral um *data mart* **focaliza uma única área de interesse** (por exemplo vendas e marketing), podendo ser montado com mais rapidez e menor custo.

Data Lake

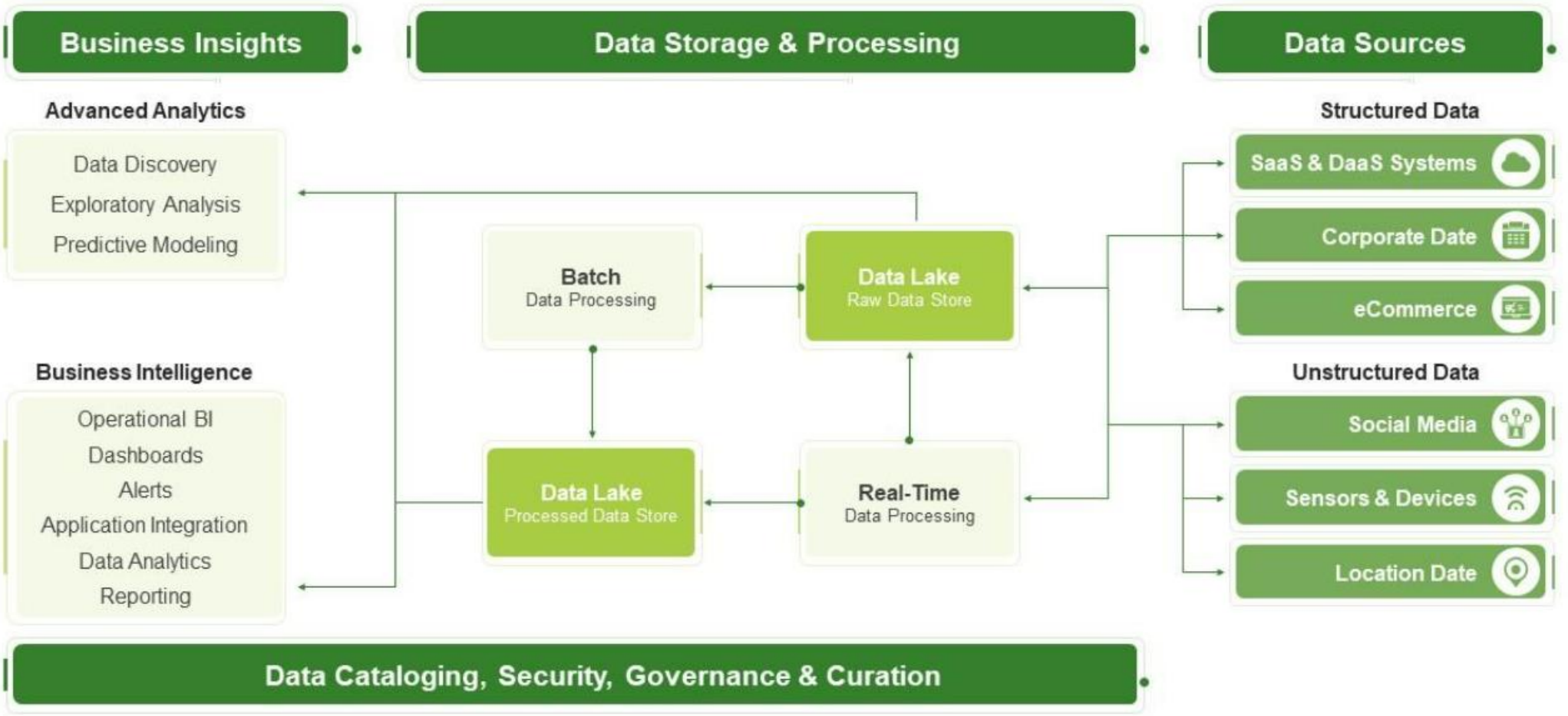
- Visão Geral Repositório de Data Lake Centralizado

Data Lake


Ingestão de dados estruturados e não estruturados
Armazenamento, segurança e proteção de dados em escala ilimitada
Catálogos e índices para análise sem movimentação de dados
Conecta dados com ferramentas de analytics e machine learning
Principais componentes de um Data Lake

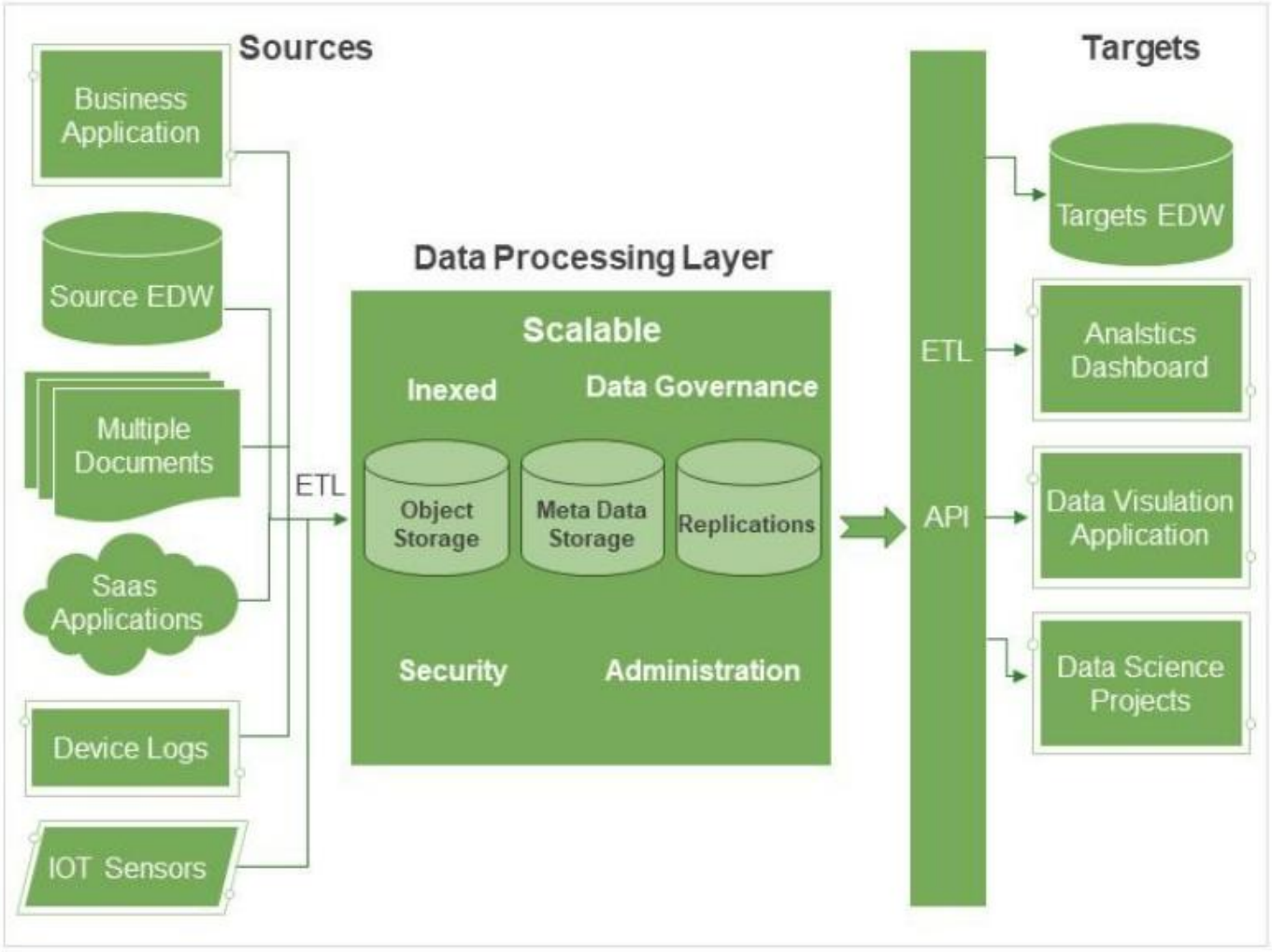
Data Lake

- Funcionamento do Repositório de Data Lake Centralizado



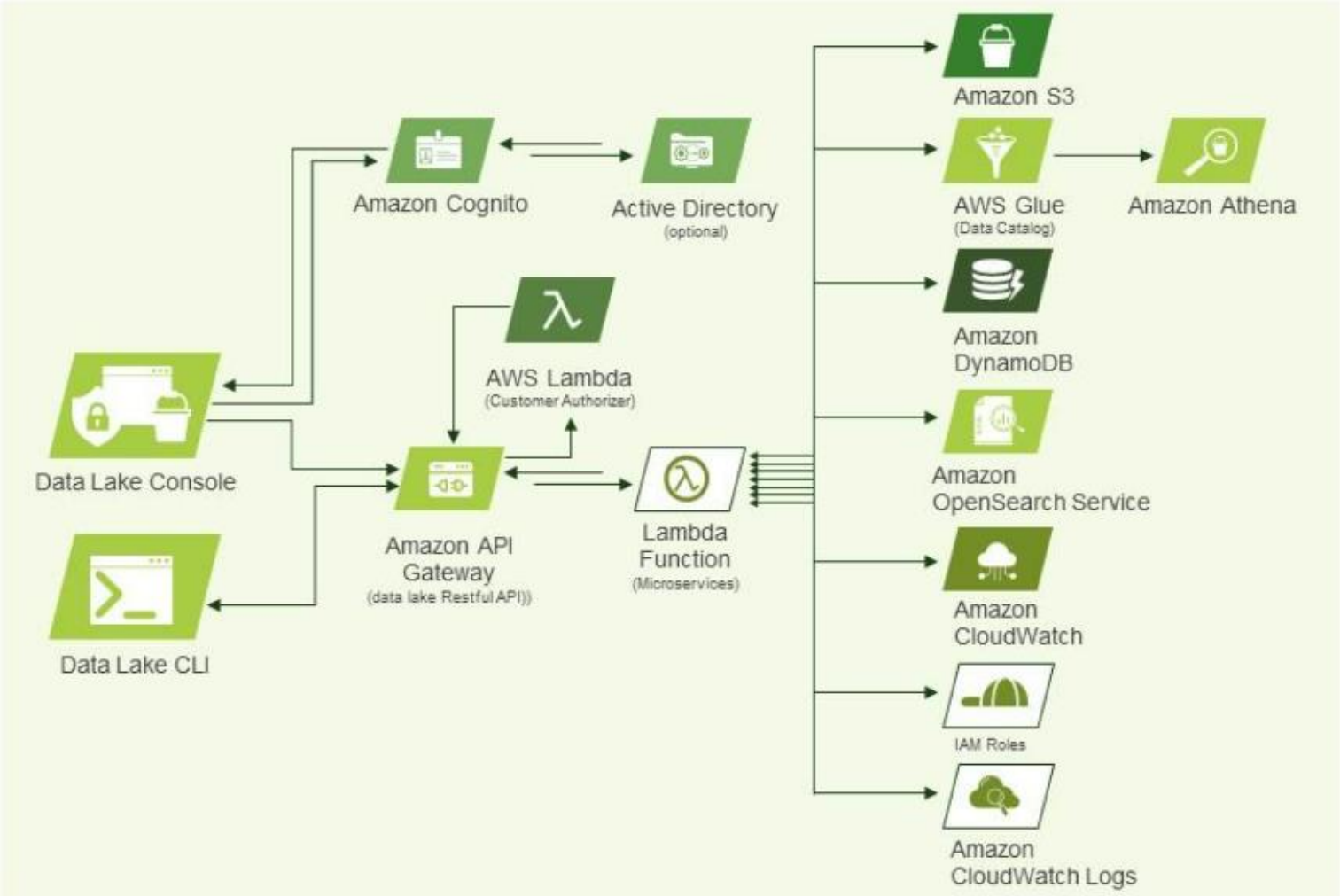
Data Lake

- Arquitetura de um Repositório de Data Lake Centralizado



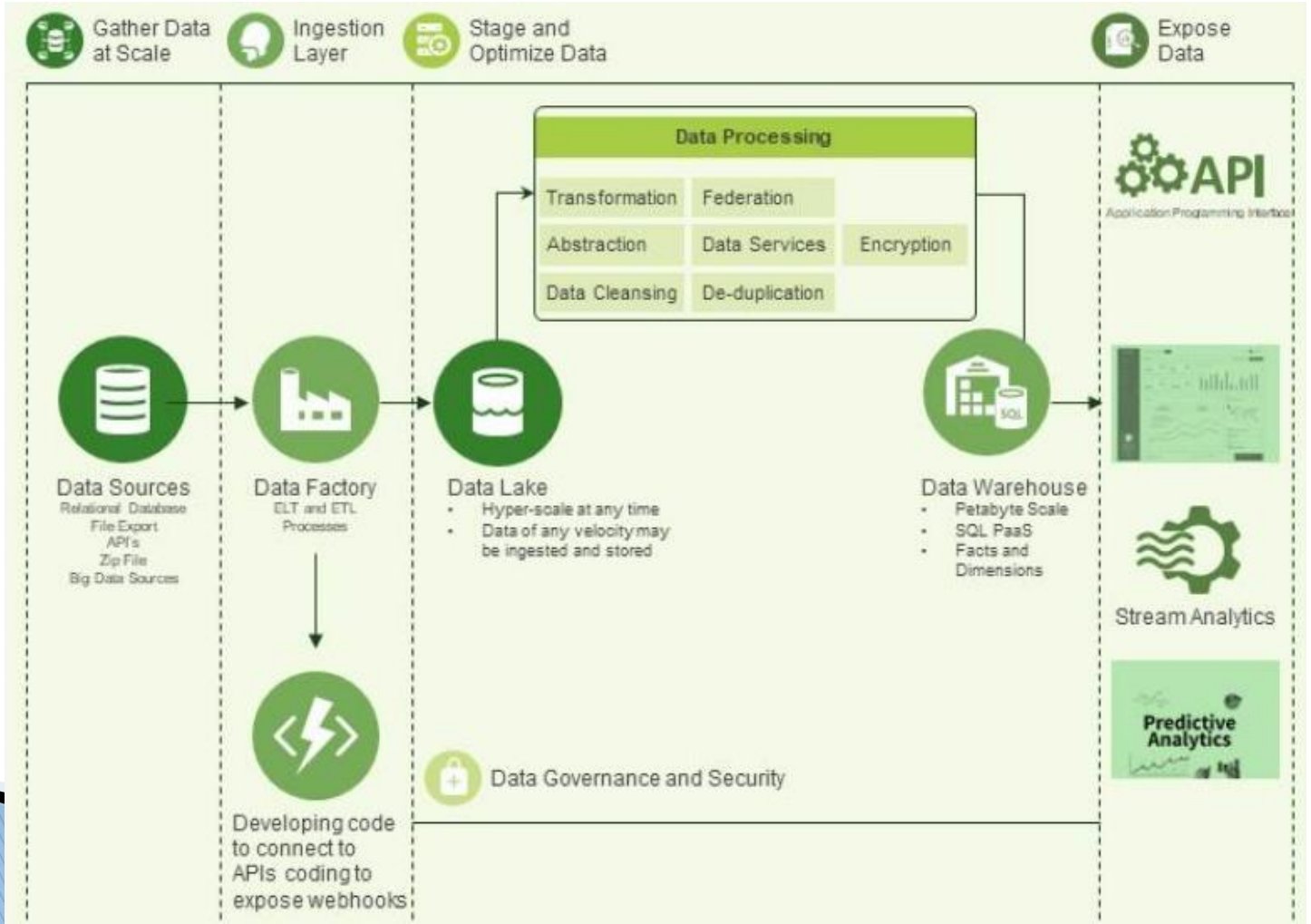
Data Lake

- Arquitetura AWS de um Repositório de Data Lake



Data Lake

- Arquitetura Azure de um Repositório de Data Lake



Data Lake

- Fornecedores Proeminentes de Repositórios de Data Lake

AWS

- Aside from Amazon EMR and S3, it provides supporting tools such as AWS Lake Formation for creating data lakes and AWS Glue for information integration and preparing
- Add text here

Oracle

- Cloud-based data lake solutions comprise a Hadoop and Spark big data service, an object storage service, and a suite of data management tools
- Add text here

Google

- Complements Dataproc and Google Cloud Storage with Google Cloud Data Fusion for data integration and a suite of services for migrating on-premises data lakes to the cloud
- Add text here

Cloudera

- May be installed in public clouds or hybrid clouds that incorporate on-premises systems, and a data lake service backs it up
- Add text here

Microsoft

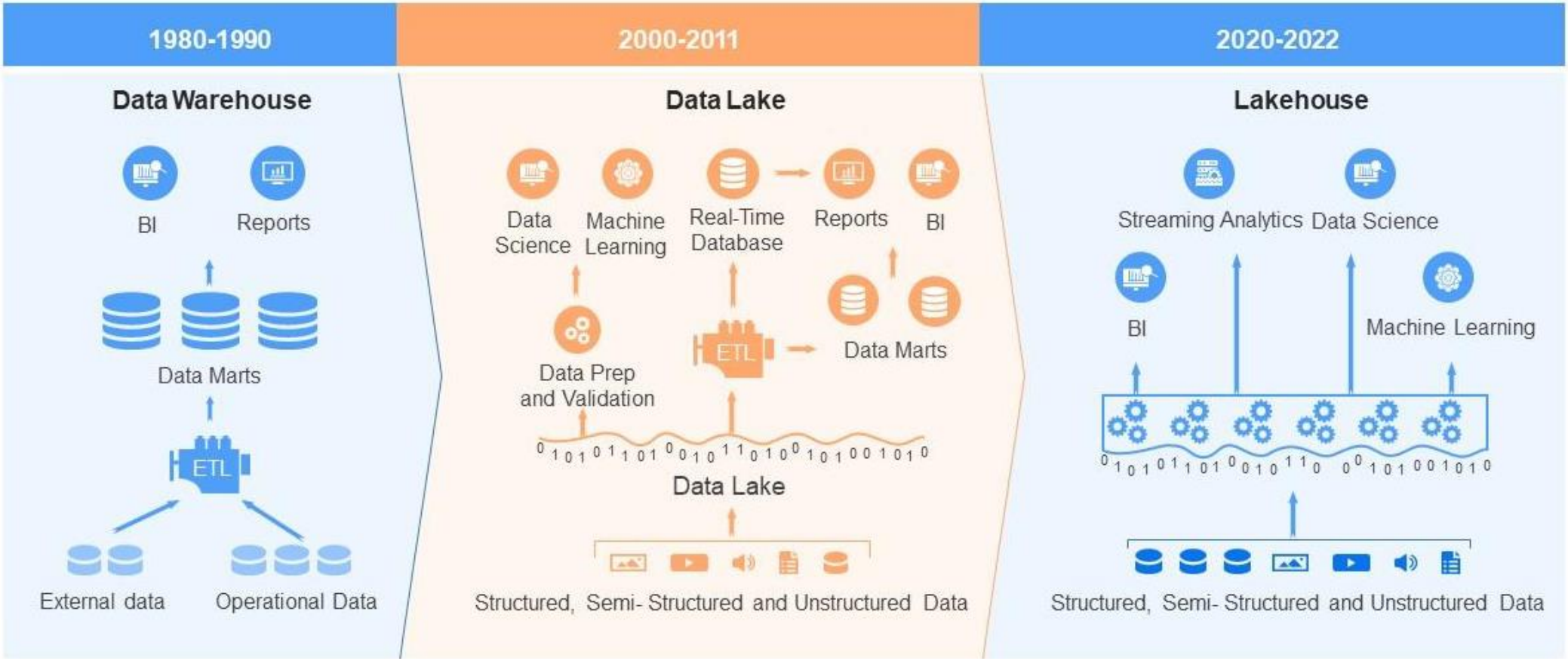
- Provide Azure Data Lake Storage Gen2, a depository that extends Blob Storage with a structured namespace along with Azure HD Insight and Azure Blob Storage
- Add text here

Snowflake

- While the Snowflake platform is primarily recognized as a cloud data warehouse vendor, it also supports data lakes and interacts with the information in cloud object stores
- Add text here

Data Lake - Benefícios

Data Lake - Comparação



Data Lake - Comparação

Factors	Data Lake	Data Lakehouse	Data Warehouse
› Types of Data	› Structured, Semi-structured and Unstructured Data	› Structured, Semi-structured and Unstructured Data	› Structured Data Only
› Cost	› $	› $	› $$$
› Format	› Open Format	› Open Format	› Closed, Proprietary Format
› Scalability	› Scales to store any amount and any type of information at low cost	› Scales to store any amount and any type of information at low cost	› Due to vendor expenses, scaling up gets massively more expensive
› Reliability	› Low quality, Data swamp	› High quality, reliable information	› High quality, reliable information
› Performance	› Poor	› High	› High
› Indented Users	› Limited: Data Scientists	› Unified: Data analysts, data scientists, machine learning engineers	› Limited: Data Analysts
› Add Text Here	› Add Text Here	› Add Text Here	› Add Text Here

Data Lake - Comparação

Data Lake – Relatório Dashboard