centro universitário FEI

AULA 16 – DATA WAREHOUSE

PROFA. DRA. LEILA BERGAMASCO

CC5232 – Banco de Dados

centro universitário FEI

NA AULA DE HOJE

- Data Warehouse
- Data Lakes

O CENÁRIO

- Corporações
  - Necessitam de decisões rápidas e precisas
  - Reação rápida a mudanças do ambiente
  - Obtenção de vantagem competitiva

- Dados
  - Disponíveis em sistemas não integrados
  - Espalhados em múltiplas e independentes plataformas
  - Dificuldade de análise

centro universitário FEI

CONCEITOS

- Processamento Operacional (OLTP)
  - Funcionalidades do negócio
  - Processamento de transações: inserção, atualização, consulta e deleção
  - Reflete valor corrente, não-redundante e atualizável
  - Altamente voláteis
  - Modelagem E/R

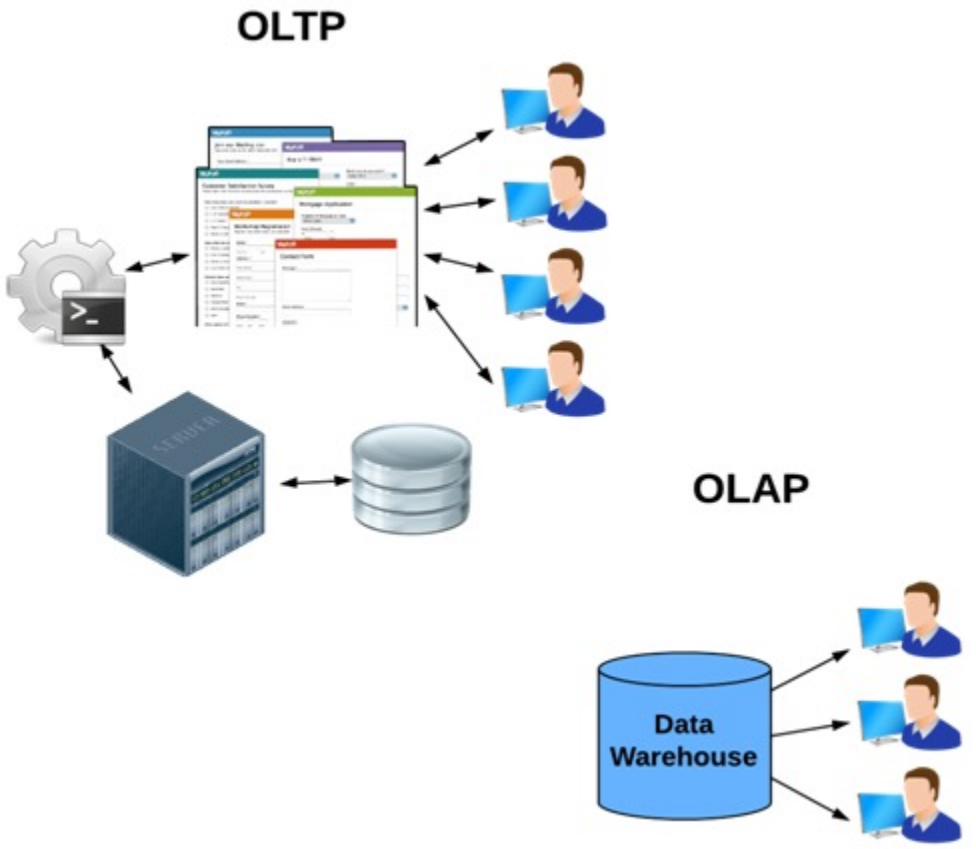
CONCEITOS

- Processamento Analítico (OLAP)
  - Suporte à tomada de decisão
  - Dados históricos, não voláteis, ready-only
  - Integram informações de diversos sistemas operacionais
  - Permitem identificações de perfis, tendências e padrões
  - Redundância de dados aceita
  - Alto desempenho na recuperação de dados vs. economia de espaço
  - Banco de Dados Multidimensional

CONCEITOS

- OLTP X OLAP

OLTP	OLAP
Orientados a aplicações	Orientados a assuntos
As Vezes de Grande tamanho	Quase sempre grandes
Dados granulados	Dados constituídos de sumarizações
Dados de pouca fontes	Dados de múltiplas fontes
Suporta consultas e atualizações	Atualizações em modo *batch*
Dados que mudam constantemente	Dados mais estáveis
Dados atuais	Dados históricos



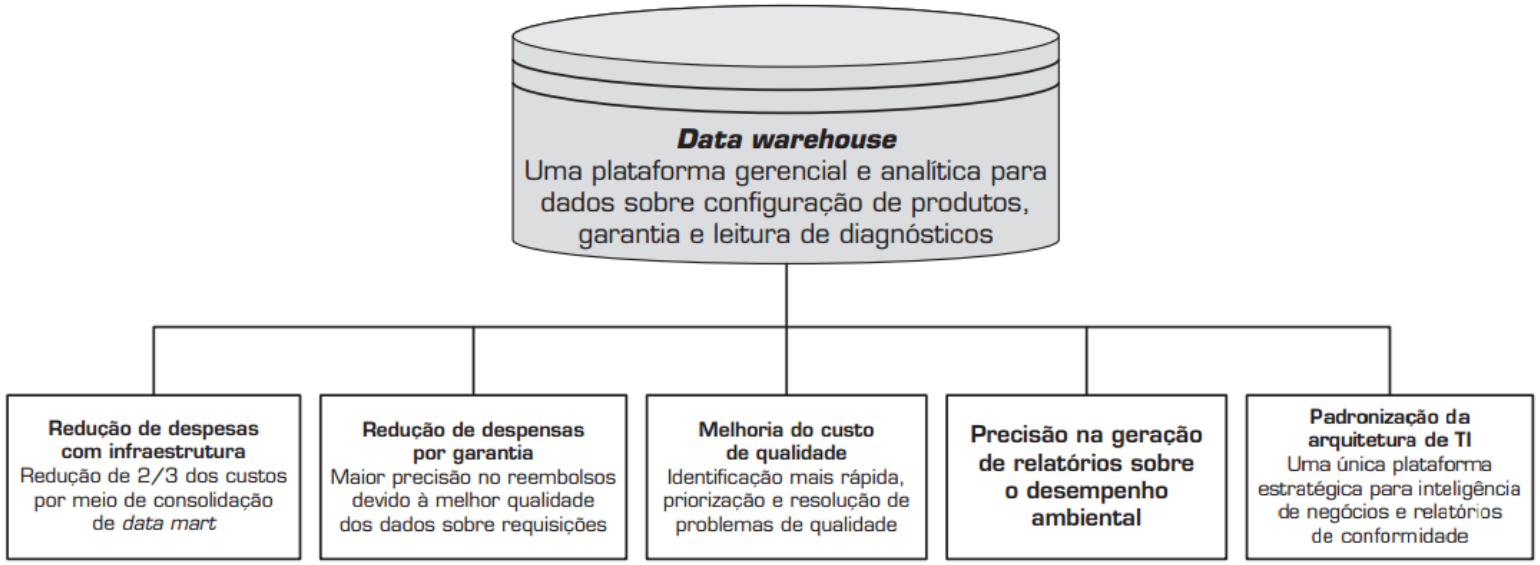
centro universitário FEI

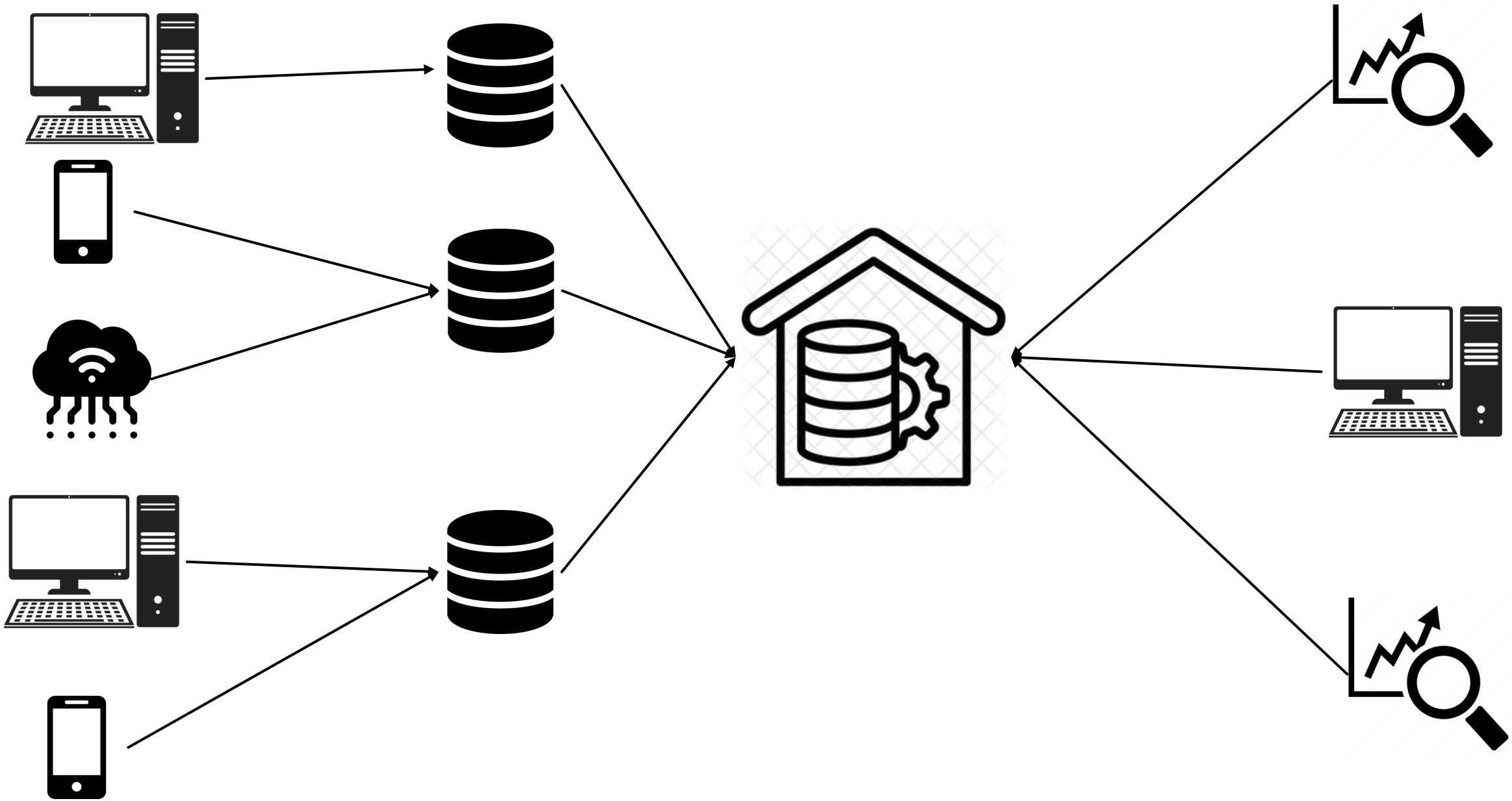
CONCEITOS

- Sistema de Apoio à Decisão (SAD)
  - Realizam processamento analítico
  - Provêm as informações necessárias ao usuário
  - Permitem análise de situações e tomada de decisões
  - Necessidades estratégicas e táticas

DATA WAREHOUSE

- Data Warehouse
  - Fornece informações para auxiliar a tomada de decisões estratégicas
  - Une, de forma organizada, informações espalhadas em diversas fontes





DEFINIÇÃO DE DW

- Data Warehouse
  - Data Warehouse é uma coleção de dados orientados à assunto, integrada, dinâmica e não-volátil, para o suporte a decisões de gerenciamento.
  - Focado na recuperação da informação.

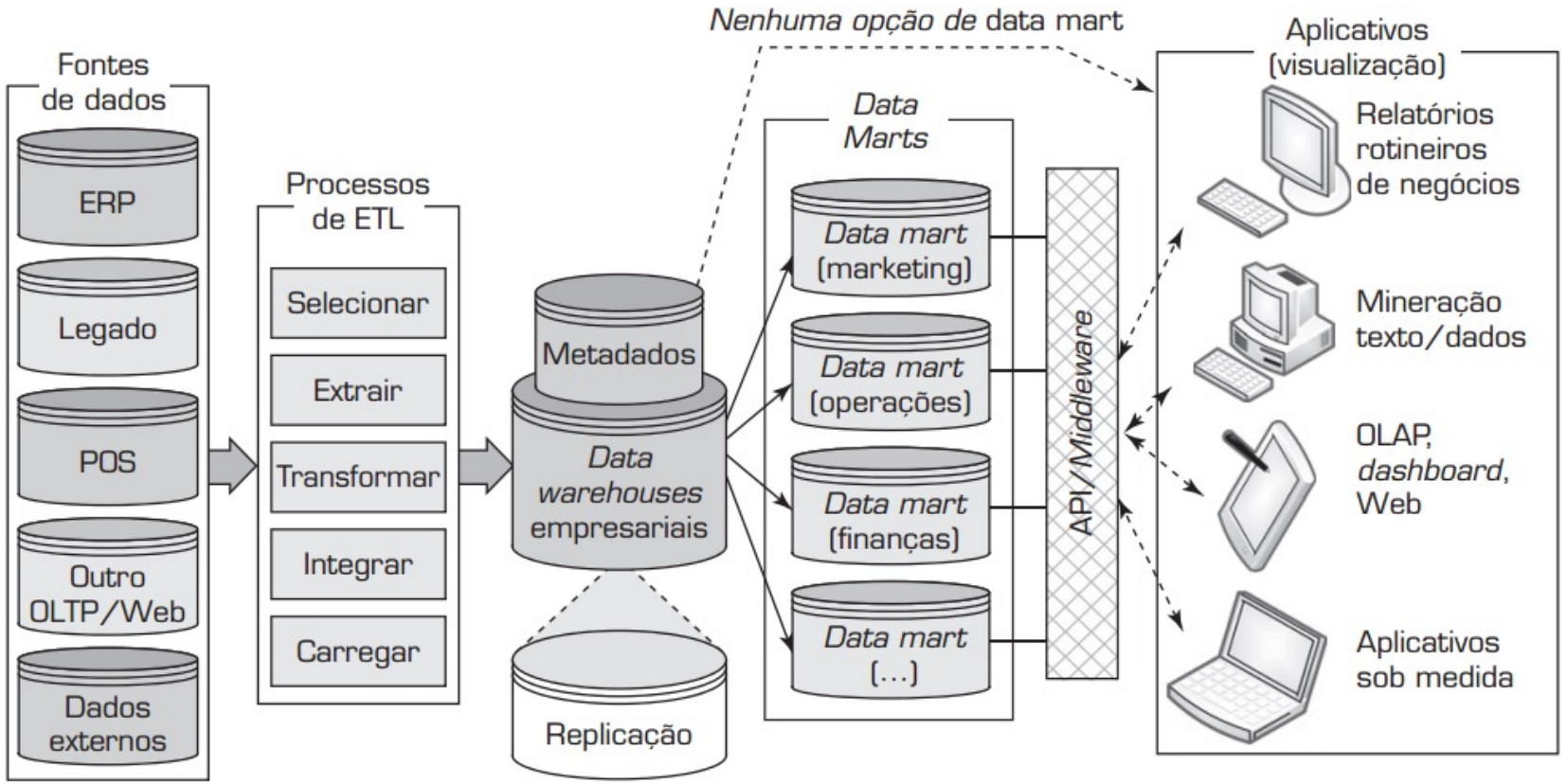
- Data Mart
  - Subconjunto lógico do DW
  - Projetado para representar uma função particular do negócio
  - Rapidamente implementável e de baixo custo
  - Controle local, em vez de centralizado
  - Redução do tempo de resposta a consultas

centro universitário FEI

DATA MART

- Problemas
  - Pode acarretar a fragmentação de dados da organização
- Solução
  - Deve haver planejamento para futura integração com um DW único de toda empresa
  - Construção de um DW na forma de DM distribuídos em unidades individuais

DATA WAREHOUSE X DATA MARTS



centro universitário FEI

CARACTERÍSTICAS BÁSICAS DE UM DW

- Orientado por tema
- Integrado
- Não-volátil
- Variante no tempo
- Dados sumarizados
- Metadados
- Dados oriundos de fontes internas e/ou externas

ORIENTADO POR TEMAS

- Refere-se ao fato do DW armazenar informações sobre temas específicos importantes para o negócio da empresa
  - Exemplos produtos, atividades, contas, clientes, etc.
- O ambiente operacional é organizado por aplicações funcionais
  - Exemplo, em uma organização bancária, estas aplicações incluem empréstimos, investimentos e seguros.

INTEGRADO

- Refere-se à consistência de nomes das unidades das variáveis
- Dados foram transformados até um estado uniforme
  - Por exemplo, todas as medidas (cm, polegadas,jardas) são convertidas para metros.

NÃO VOLÁTIL

- Permite o "load-and-access”
- Os dados após serem extraídos, transformados e transportados para o DW estão disponíveis aos usuários somente para consulta

VARIANTE NO TEMPO

- Os DW armazenam dados por um período de tempo de 5 a 10 anos
- Refere-se a algum momento específico
  - não é atualizável
- No DW haverá sempre uma tabela dimensão ou fato, cuja estrutura registrará o elemento tempo

centro universitário FEI

METADADOS

- “Dados sobre dados”
- Provêm informações sobre a estrutura de dados e as relações entre estas dentro ou entre bancos de dados
- “São todas as informações do ambiente do DW que não são seus próprios dados”

GRANULARIDADE

- É o nível de detalhes dentro do banco de dados do DW
- Quanto menor a granularidade, maior o nível de detalhes e, consequentemente, maior o volume de dados armazenado
- Exemplo, Registro de Vendas de uma rede de supermercados:
  - diária: sumarização de vendas e carga diária no Banco de Dados
  - mensal: sumarização de dados e carga a cada 30 dias no Banco de Dados

AGREGAÇÃO

- São registros sumarizados logicamente redundantes com os dados básicos do DW

VENDAS - LOJA XX (R$)					
FILIAL	DIA				
	13/09	14/09	15/09	16/09	17/09
1	3000	2500	2000	3000	5000
2	1000	1500	1200	1800	2500
3	4000	3500	3000	3400	6000

VENDAS - LOJA XX (R$)	
FILIAL	SEMANA
1	15500
2	8000
3	19900

- Finalidades:
  - melhorar o tempo de reposta as consultas
  - reduzir o tempo de processamento
  - reduzir espaço de armazenamento

MODELAGEM DE UM DW

- Problemas da Modelagem E/R
  - Redução de visão global do negócio para grandes modelos
  - Não tem alto desempenho na recuperação de dados (principalmente joins)
  - Para cada variação na estrutura do modelo, há necessidade de reescrever e ajustar as implementações

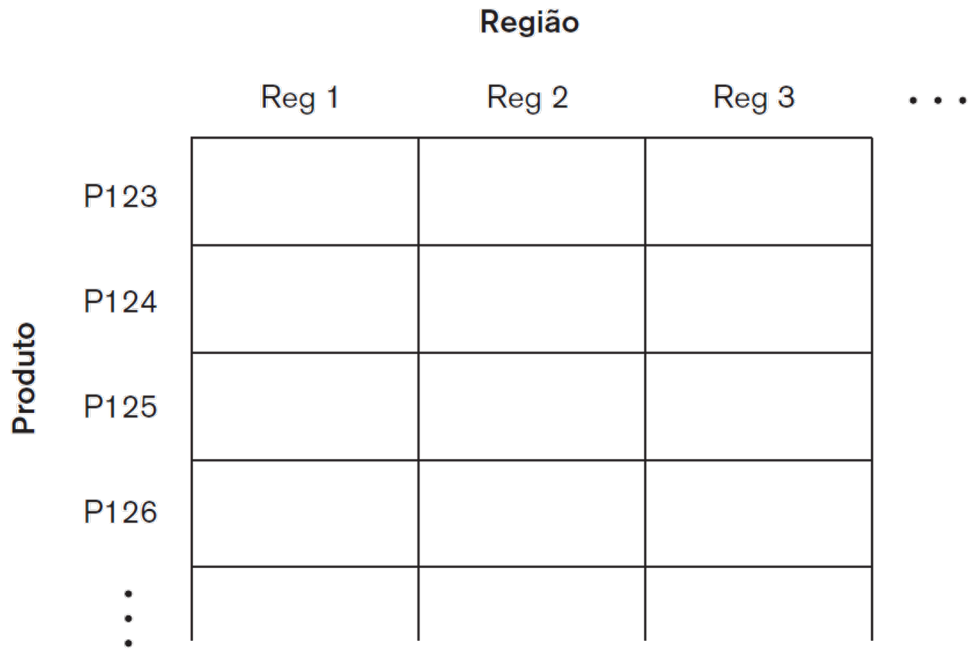
MODELAGEM DE UM DW

- Modelagem Dimensional
  - Específica para processamento analítico
  - Apresentação de dados padronizada, intuitiva e que permite alto desempenho de acesso
  - Dois tipo de tabelas: Fato e dimensão.
  - Chave primária simples da tabela dimensão corresponde à chave estrangeira de fato (Esquema estrela)

MODELAGEM DE UM DW

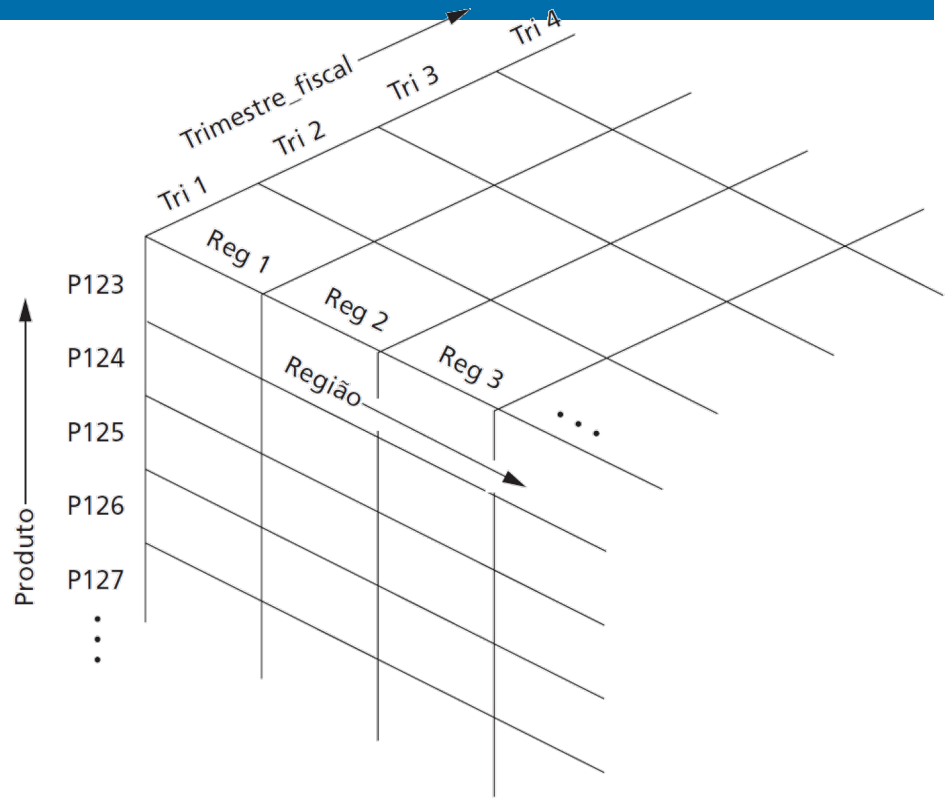
- Uma tabela é bidimensional
- Problema de vendas:
  - Você é um fabricante de produtos hospitalares e precisa armazenar os dados de venda em diferentes regiões do Estado de São Paulo

E se eu quisesse saber ao longo de diferentes trimestres?



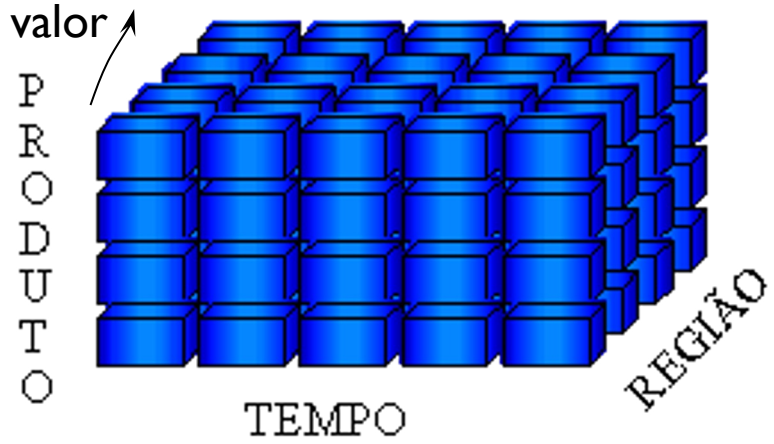
MODELAGEM DE UM DW

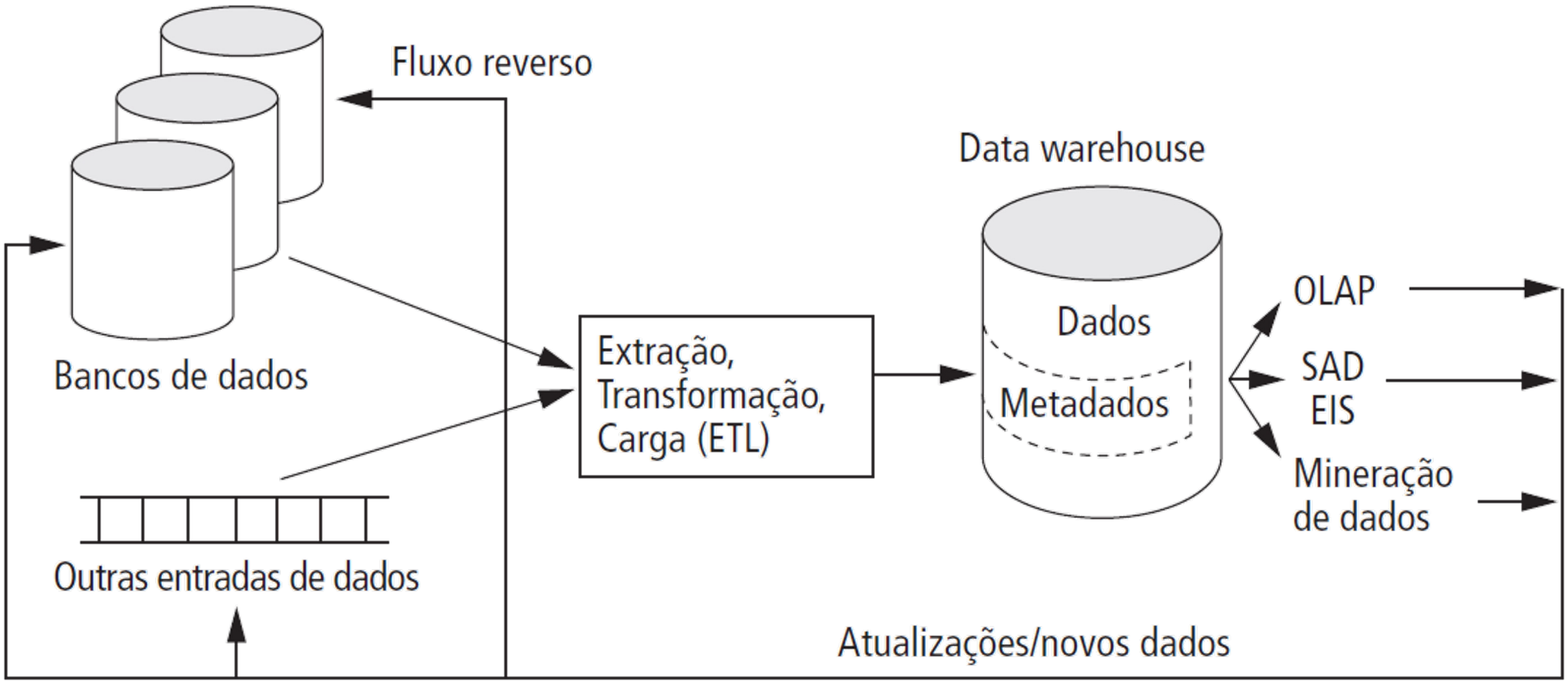
- Modelos multidimensionais tiram proveito dos relacionamentos inerentes nos dados para preencher os dados em matrizes multidimensionais, chamadas cubos de dados.
- Ao acrescentar uma dimensão de tempo, como os trimestres fiscais de uma organização, seria produzida uma matriz tridimensional, que poderia ser representada usando um cubo de dados:



MODELAGEM DIMENSIONAL

- Também chamado de hipercubo




Fluxo reverso
Data warehouse
Bancos de dados
Extração,
Transformação,
Carga (ETL)
Dados
Metadados
OLAP
SAD
EIS
Mineração
de dados
Outras entradas de dados
Atualizações/novos dados

O AMBIENTE DE UM DW

- Arquitetura resumida de DW

BACK END:
- extração de dados de sistemas legados e fontes externas
- consolidação e sumarização de dados
-carga dentro do warehouse

FRONT END:
- usuário acessa o warehouse
- análise de dados

MODELAGEM DIMENSIONAL

- O modelo multidimensional (também chamado de modelo dimensional) envolve dois tipos de tabelas: tabelas de dimensão e tabelas de fatos:

- Tabelas fatos
  - Contêm as medições numéricas do negócio
    - Exemplo: unidades_vendidas, custo_dolar
  - Grande quantidade de dados
  - Chave primária composta por FKs
  - Atributos numéricos e valorados

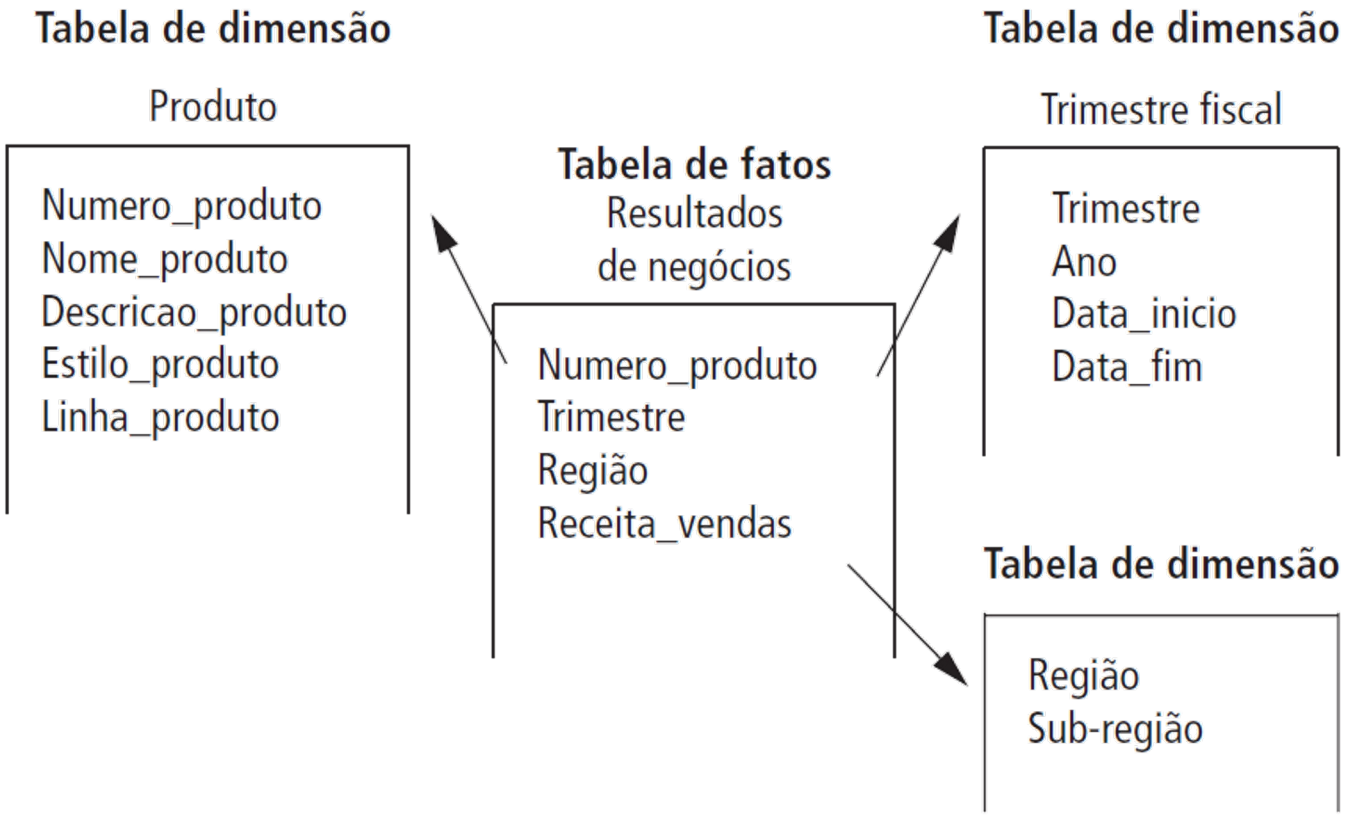
centro universitário FEI

MODELAGEM DIMENSIONAL

- Tabelas dimensão
  - Contém dados descritivos do negócio
  - Chave primária simples
  - Pequena quantidade de informações se comparadas com as tabelas fato
  - Modelos reais contêm entre 4 e 15 dimensões
  - Modelos com mais de 20 dimensões devem ser melhor estudados

ESQUEMA ESTRELA

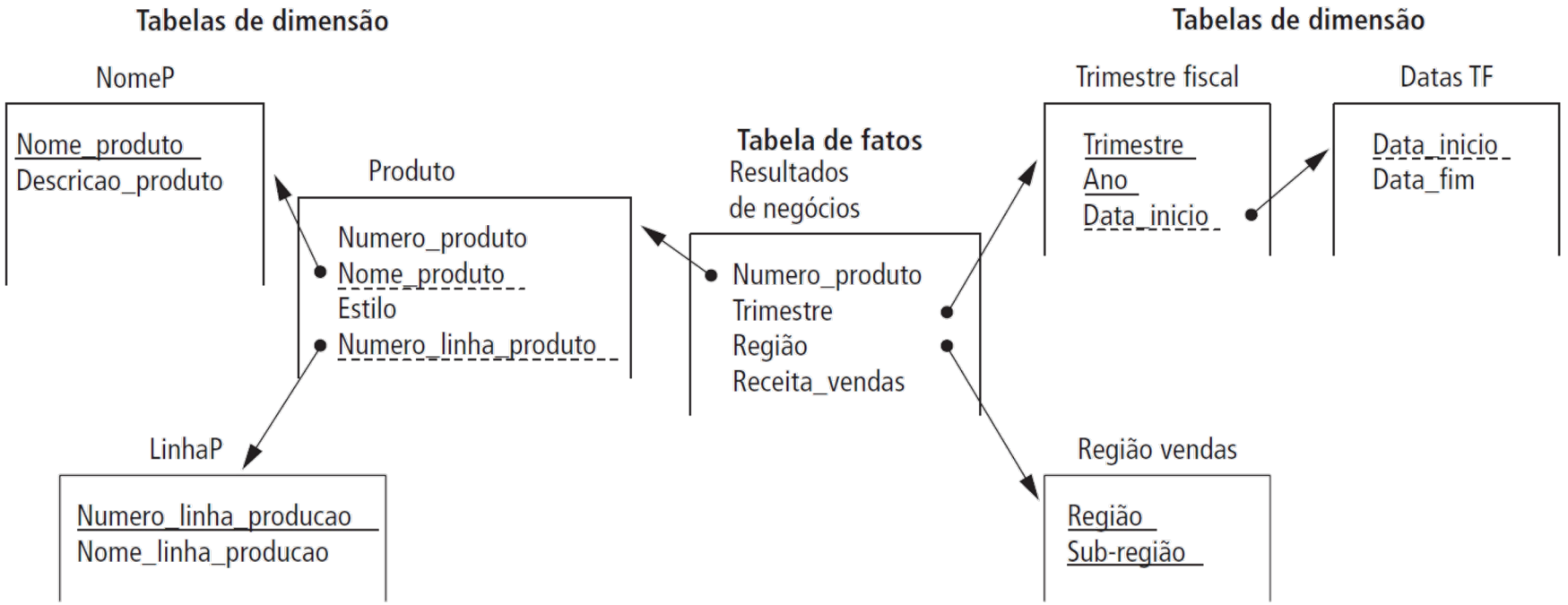
- Este esquema é chamado de estrela, por apresentar a tabela de fatos "dominante" no centro do esquema e as tabelas de dimensões nas extremidades.



Quem é chave primária e chave estrangeira?

ESQUEMA FLOCO DE NEVE

- O esquema floco de neve é uma variação do esquema estrela em que as tabelas dimensionais de um esquema estrela são organizadas em uma hierarquia :



centro universitário FEI

QUAL MELHOR?

- Floco de neve:
  - Normalização alta
  - Problemas na recuperação
- Estrela:
  - Mais simples
  - Mais dispendioso no espaço

TOPOLOGIAS DE DWS

- Centralizada
  - Único Banco de Dados Físico
  - usados onde existe uma necessidade comum de informações

- Distribuída
  - Vários DW interligados através de uma rede com forte suporte a processamento distribuído
  - Usuário pode conectar-se a qualquer DW
  - Apresenta problemas de desempenho
  - Será muito utilizada para dar suporte às aplicações para Web.

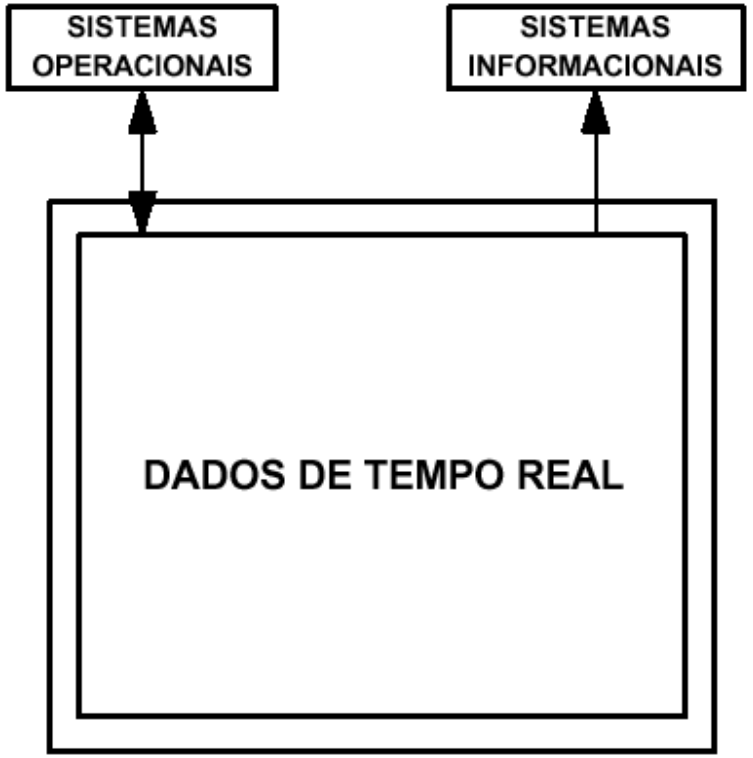
centro universitário FEI

ARQUITETURA DE UM DW

- Arquitetura de Dados
  - Uma camada (one tier)
    - Dados armazenados uma única vez
  - Duas camadas (two tier)
    - Dados operacionais e analíticos separados em camadas distintas
  - Três camadas (three tier)
    - Transformação de dados não é executada em um único passo

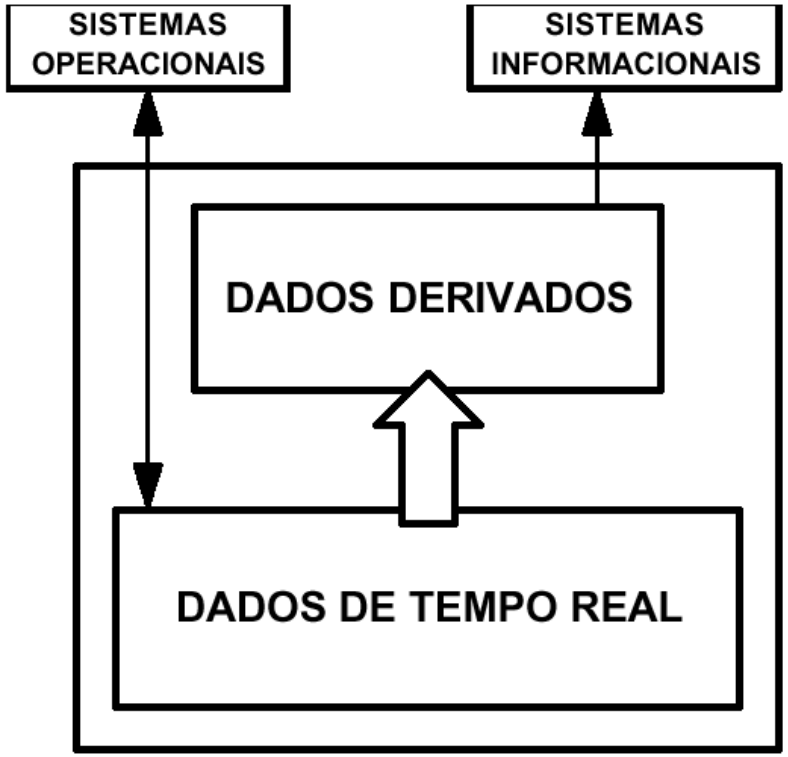
ARQUITETURA DE DADOS DO DW

- Uma camada



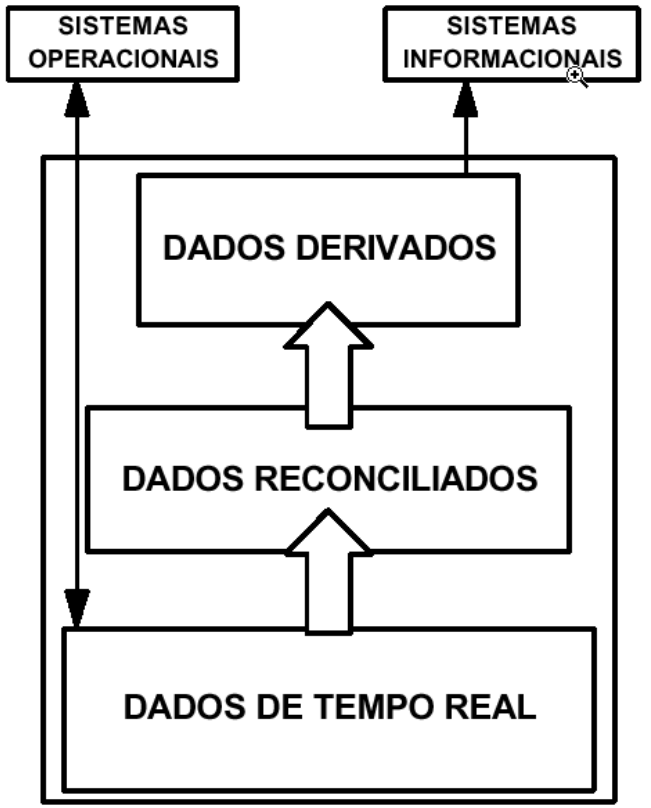
ARQUITETURA DE DADOS DO DW

- Duas camadas



ARQUITETURA DE DADOS DO DW

- Três camadas



COMO CRIAR UM DW

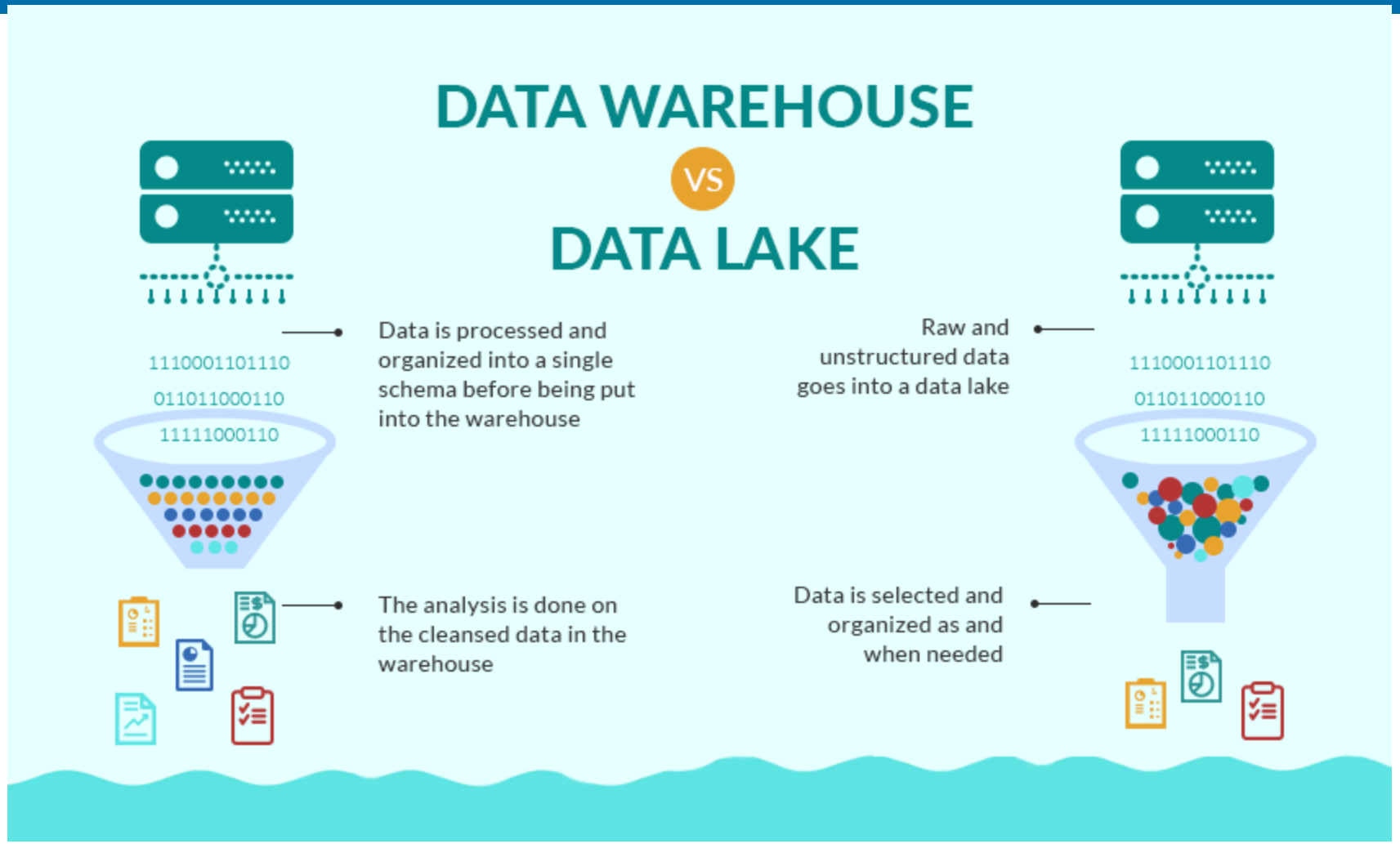
- A aquisição de dados para o data warehouse envolve as seguintes etapas:
- Os dados precisam ser extraídos de várias fontes heterogêneas.
- Os dados precisam ser formatados por coerência dentro do data warehouse.
- Os dados precisam ser limpos para garantir a validade.
- Os dados precisam ser ajustados ao modelo de dados do data warehouse.
- Os dados precisam ser carregados no data warehouse.

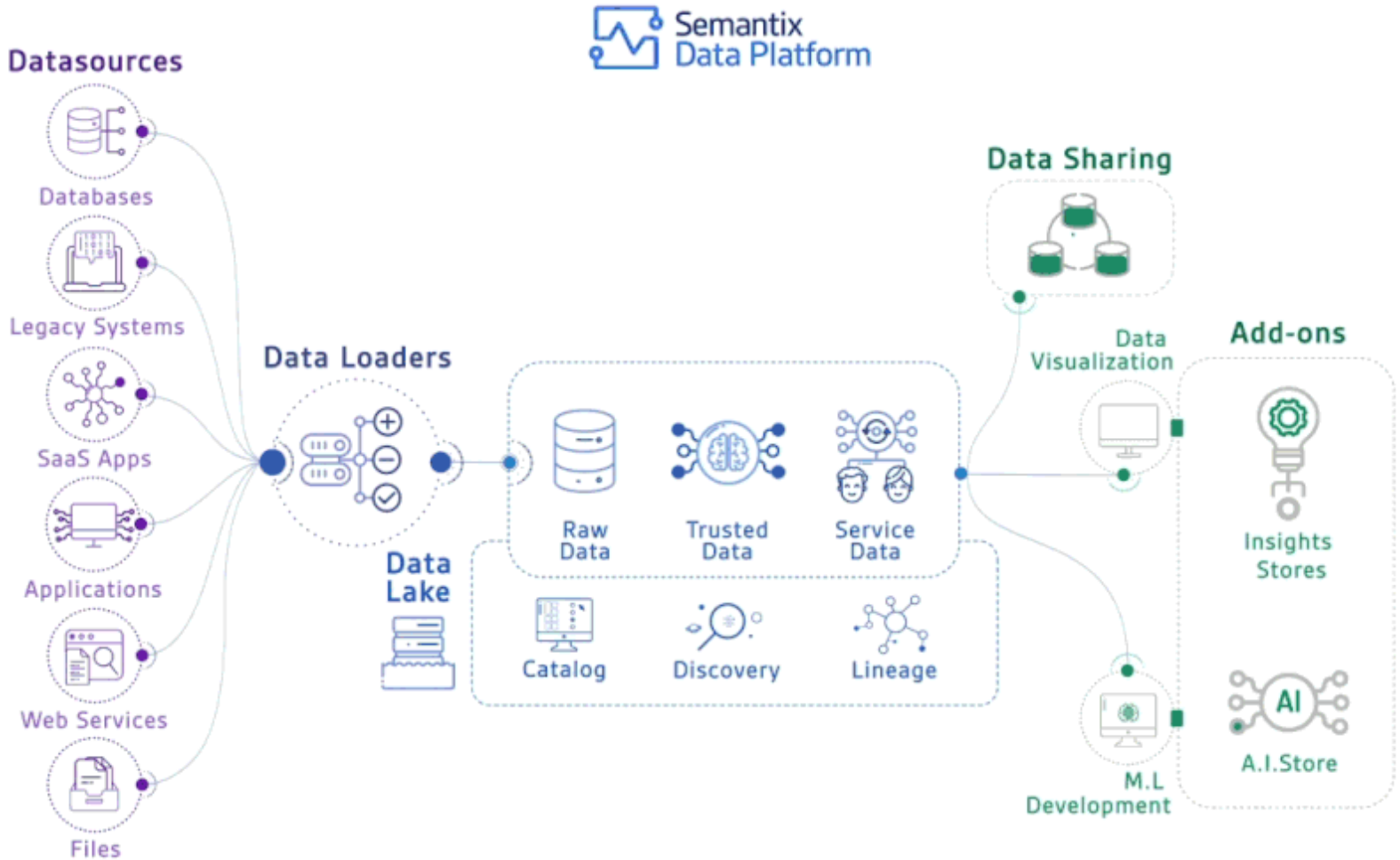
centro universitário FEI

PROBLEMAS DE UM DW

- Manutenção
  - Outras equipes irão consultar o seu DW?
- Mudanças das fontes de coleta
- Mudanças de necessidades de negócio

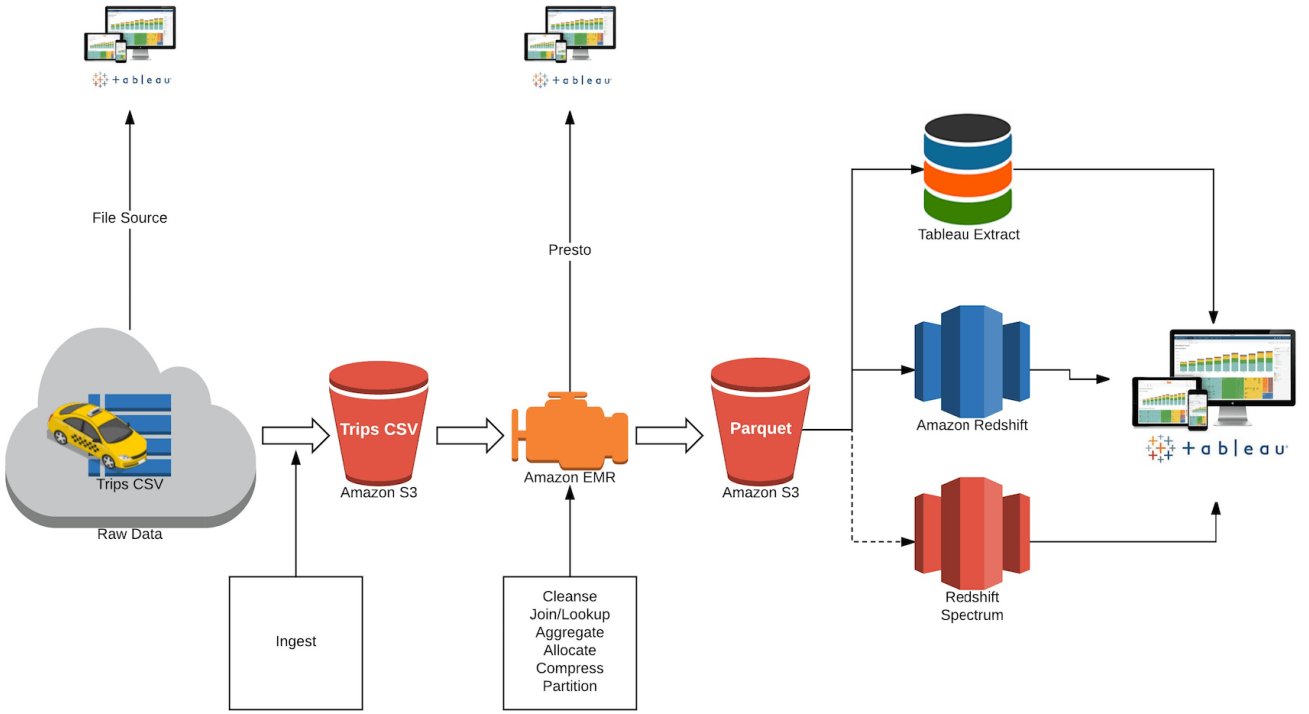
DATA WAREHOUSE X DATA LAKES

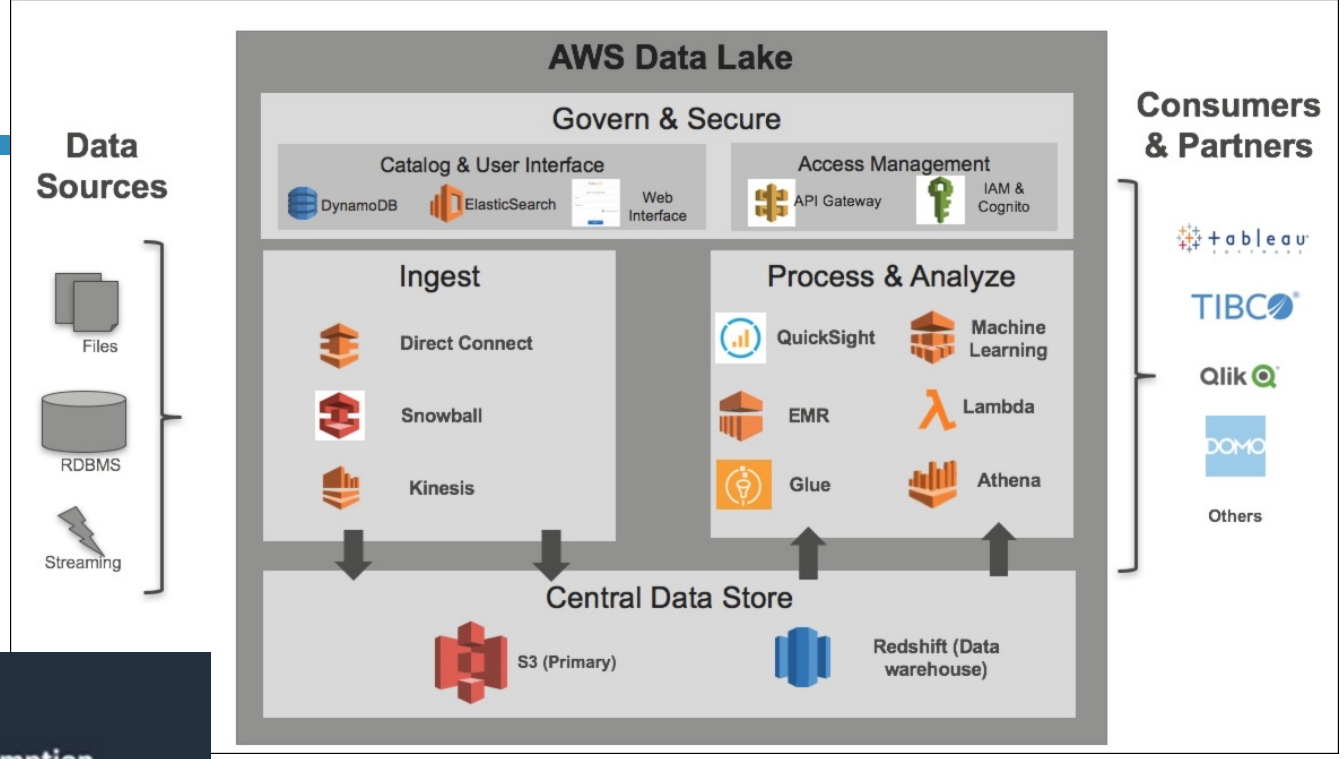


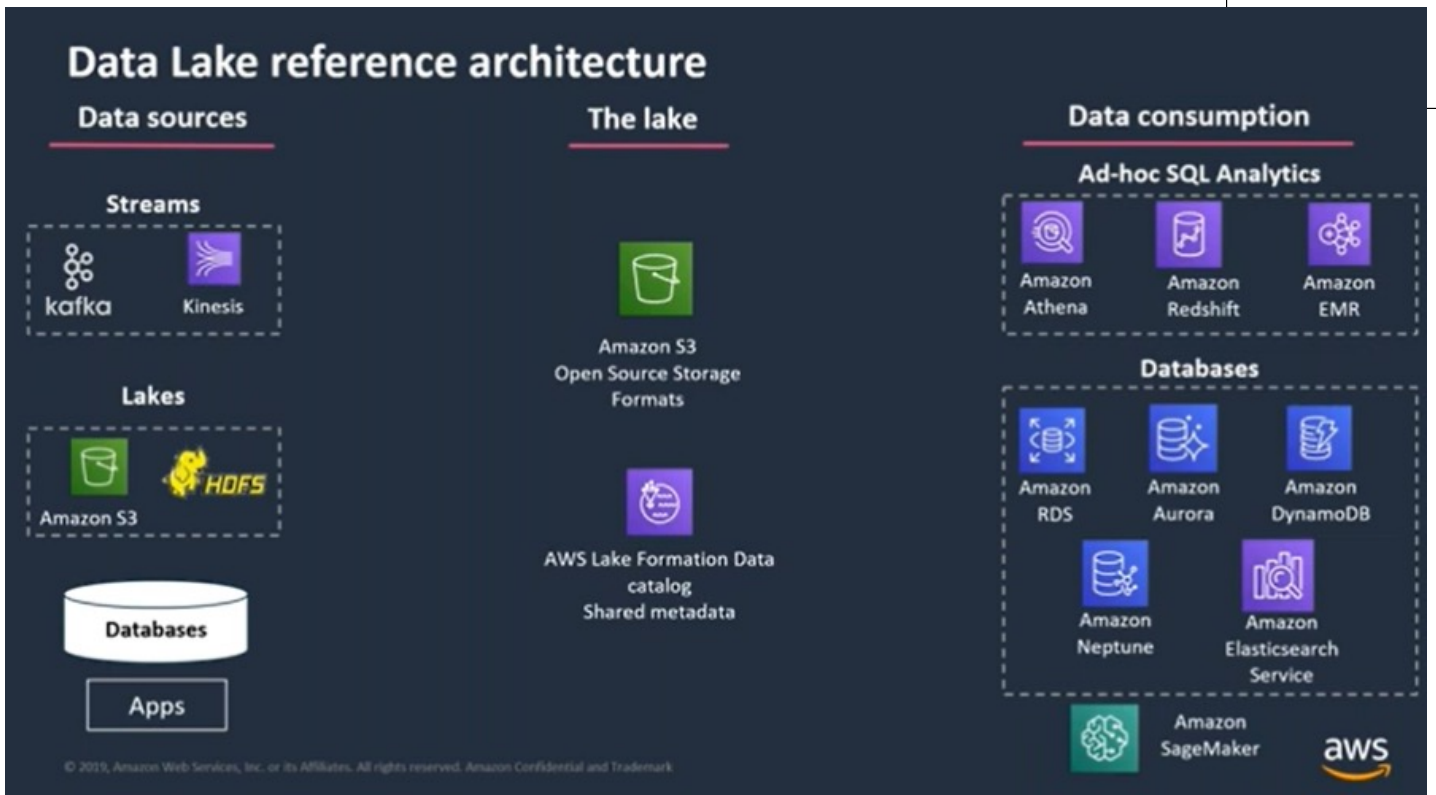

Semantix
Data Platform
Datasources
Databases
Legacy Systems
SaaS Apps
Applications
Web Services
Files
Data Loaders
Data
Lake
Raw
Data
Trusted
Data
Service
Data
Catalog
Discovery
Lineage
Data Sharing
Data
Visualization
M.L
Development
Add-ons
Insights
Stores
AI
A.I.Store

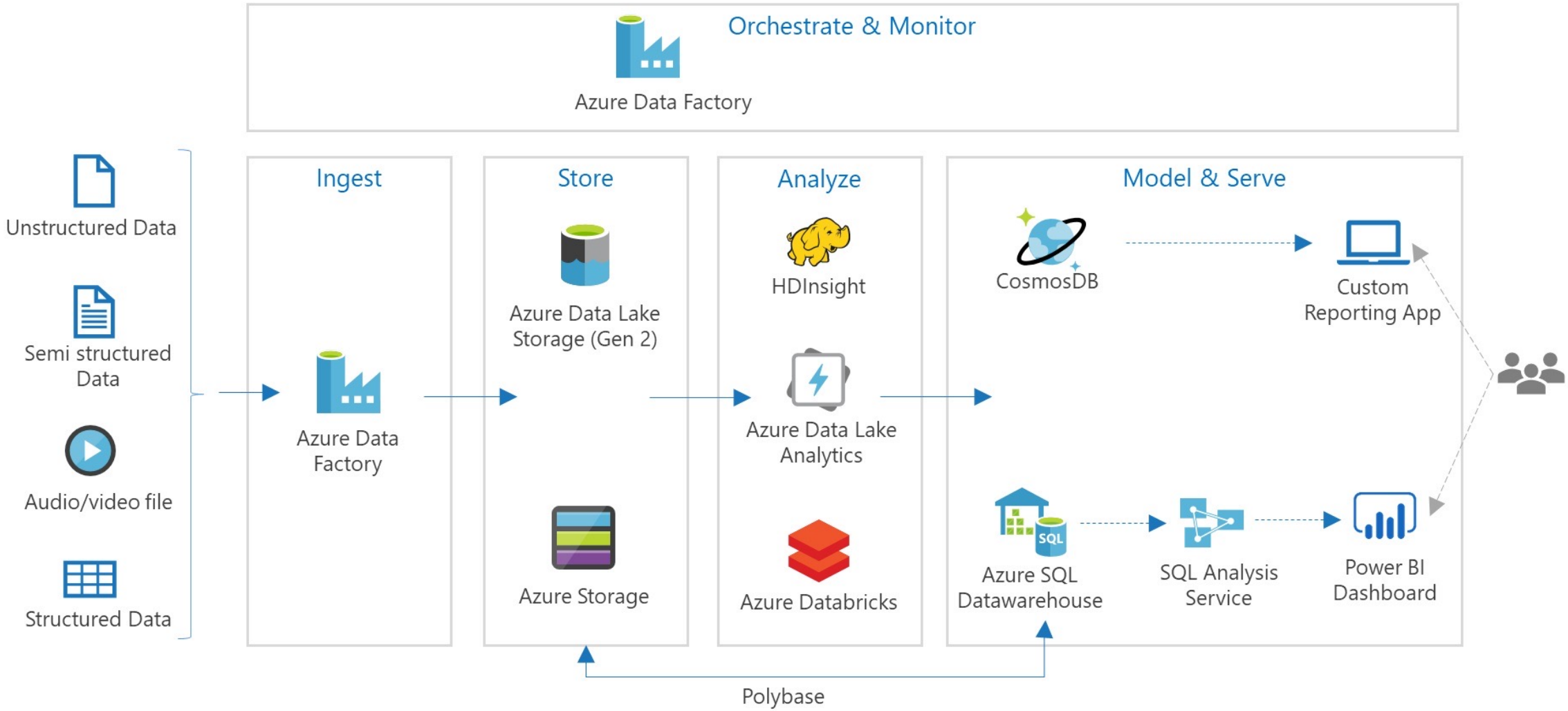
FERRAMENTAS

- DBMINER
- MS SQL Server
- AWS Redshift
- Azure DW




Data Sources
Files
RDBMS
Streaming
AWS Data Lake
Govern & Secure
Catalog & User Interface
DynamoDB
ElasticSearch
Web Interface
Access Management
API Gateway
IAM & Cognito
Ingest
Direct Connect
Snowball
Kinesis
Process & Analyze
QuickSight
Machine Learning
EMR
Lambda
Glue
Athena
Central Data Store
S3 (Primary)
Redshift (Data warehouse)
Consumers & Partners
tableau
TIBCO
Qlik
DOMO
Others


Data Lake reference architecture
Data sources
Streams
kafka
Kinesis
Lakes
Amazon S3
HDFS
Databases
Apps
The lake
Amazon S3
Open Source Storage Formats
AWS Lake Formation Data catalog
Shared metadata
Data consumption
Ad-hoc SQL Analytics
Amazon Athena
Amazon Redshift
Amazon EMR
Databases
Amazon RDS
Amazon Aurora
Amazon DynamoDB
Amazon Neptune
Amazon Elasticsearch Service
Amazon SageMaker
aws
© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved. Amazon Confidential and Trademark


Orchestrate & Monitor
Azure Data Factory
Unstructured Data
Semi structured Data
Audio/video file
Structured Data
Ingest
Azure Data Factory
Store
Azure Data Lake Storage (Gen 2)
Azure Storage
Analyze
HDInsight
Azure Data Lake Analytics
Azure Databricks
Model & Serve
CosmosDB
Custom Reporting App
Azure SQL Datawarehouse
SQL Analysis Service
Power BI Dashboard
Polybase

centro universitário FEI

CONCLUSÕES

- Data Warehouse integra grandes volumes de dados originados em sistemas separados
- Necessitam de grande esforço para seu desenvolvimento
- Torna possível a descoberta de conhecimento escondido nos dados
- Útil para organizações que precisem tomar decisões estratégicas de risco e que necessitem se posicionar de forma vantajosa
- Desenvolvimento de servidores de BD paralelos poderá viabilizar o suporte a Data Warehouses cada vez maiores
- De 2009
  - Tratará dados multimídia
  - Data Warehouse deverá também ser viabilizado na Internet

centro universitário FEI

OBRIGADO E ATÉ A PRÓXIMA AULA!