

# Construindo um pipeline de Qualidade de Dados

---

Trilha – Big Data

Tatiana Nascimento

Arquitetura e Engenharia de Dados



## THE DEVELOPER'S CONFERENCE

---

---



# Agenda

Data Warehouse e Data Lake.

Data Lake e Data Swamp (Pântano).

Objetivos de uma organização orientada a dados.

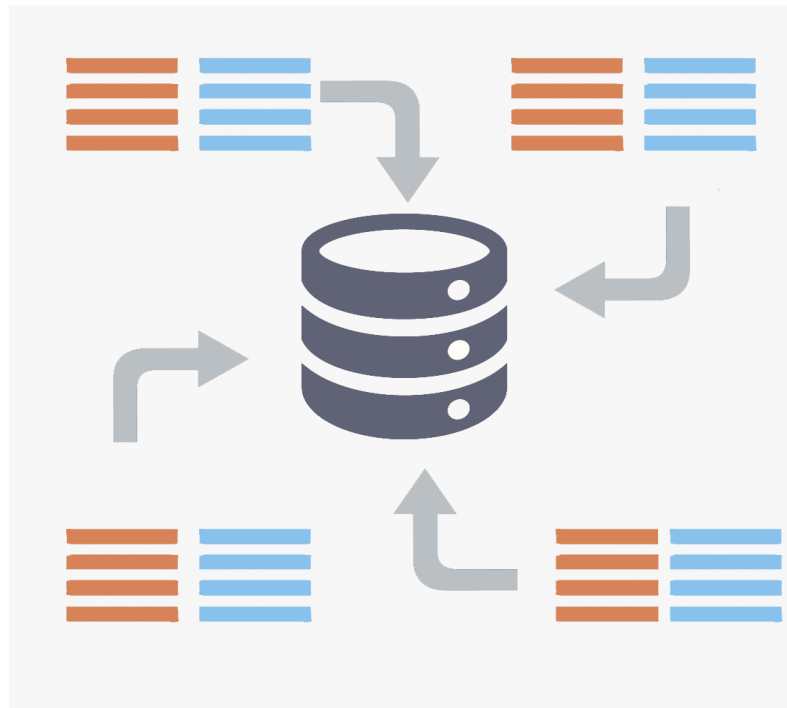
Framework Governança de Dados

Pipeline Data Quality

Ferramentas

Conclusão

# Data Warehouse e Data Lake




- Processo definido
- Ambiente único
- Ferramentas
- D -1



- Processo não tão definido
- Vários ambientes
- Várias ferramentas
- Quando relevante



- 
- Limpo
  - Atualizado
  - Conhecido
  - Acessos definidos
  - Culturas e Processos estabelecidos



# Data Swamp

- Desconhecido
- Desconfortável
- Duvidoso

Tem só um monstrinho...

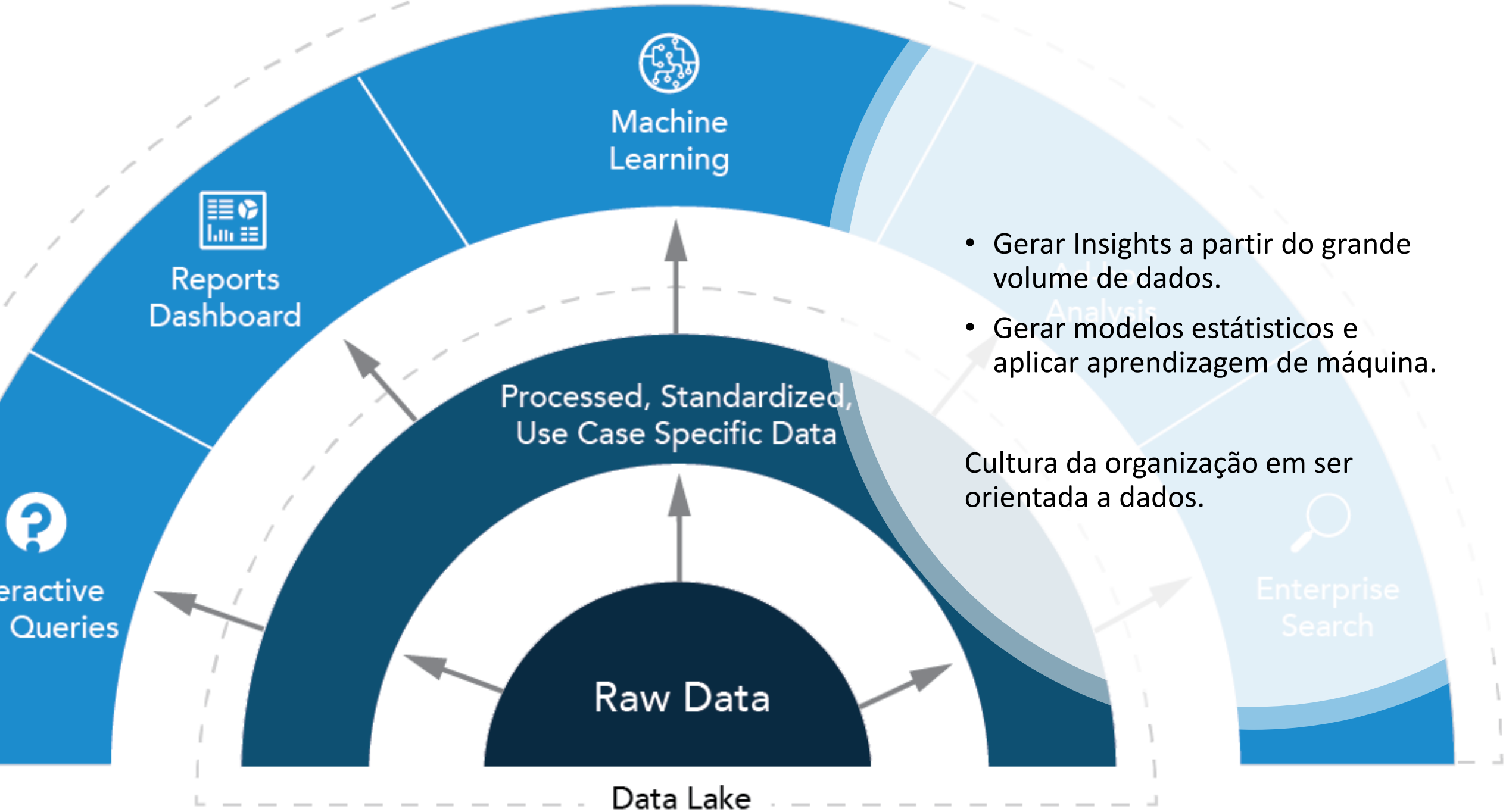






Mas pode piorar.







Framework  
DAMA -DMBOK

---





**DAMA-DMBOK2 Data Management Framework**

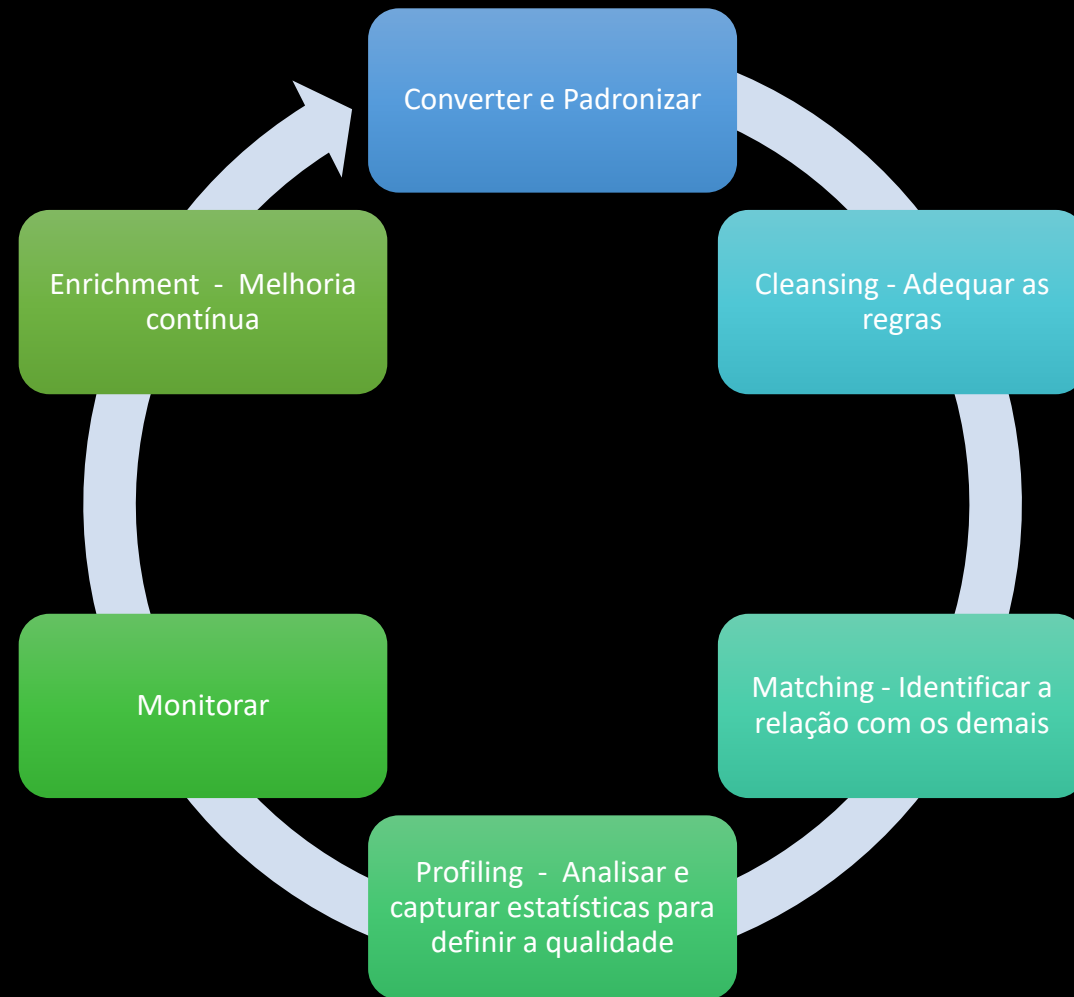
A photograph of two people in an office environment. On the left, a man with a shaved head, wearing a brown jacket, is smiling broadly and giving a high-five. On the right, a woman with dark hair, wearing a grey sweater, is also smiling and giving a high-five. They are standing in front of a wooden desk with several laptops and papers. The background shows a window with a grid pattern. A large, semi-transparent dark circle is overlaid on the right side of the image, containing text.

# Data Quality em Big Data

- Data Owner
  - Dicionário de dados
- 
- Uso de amostras
  - Monitoria dos dados
  - Streaming de Dados como estratégia.
  - Data Safe – Data Marts



# Pipeline base de qualidade de dados



# Assessment

Definir os critérios de negócios e os ranges e estatísticas aceitáveis para atributos e regras.

Monitoria e Melhoria continua

Acurância

Consistência

Relevância do dado

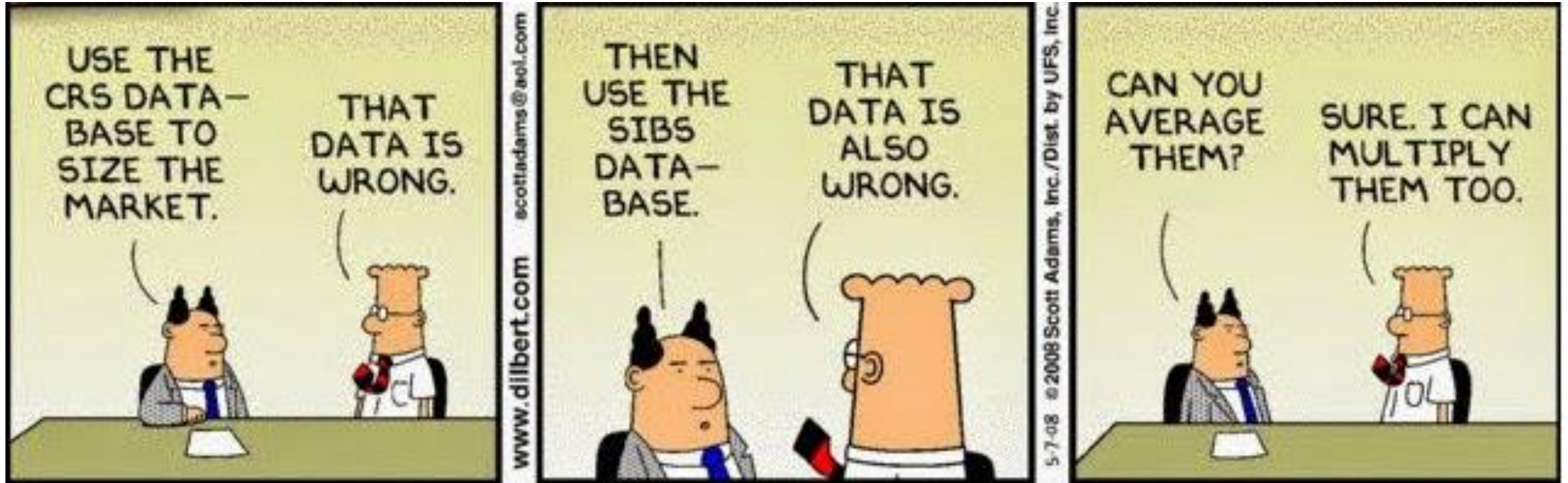
| Atributo        | Nulo | Critério de negócio<br>$X \geq 100,00$ | Aceite   | Resultado   |
|-----------------|------|--|----------|-------------|
| Nome do Cliente | 10%  | -                                      | 5% a 10% | Passou      |
| Valor do frete  |      | 90%                                    | 0%       | Inaceitável |



Figure 1. Magic Quadrant for Data Quality Tools



Source: Gartner (March 2019)



<https://www.youtube.com/watch?v=KPMNj26Z85E>





Muito  
obrigada!!!

Tatiana Nascimento

[tatianasn@gmail.com](mailto:tatianasn@gmail.com)

LinkedIn -

[/in/Tatiananascimento](#)

Anexo

---

<https://www.informatica.com/data-quality-magic-quadrant.html#fbid=oBCXUdYun2T>

---

<https://dama.org/content/body-knowledge>