Faculdade XPe

Engenheiro(a) de Dados IA Expert

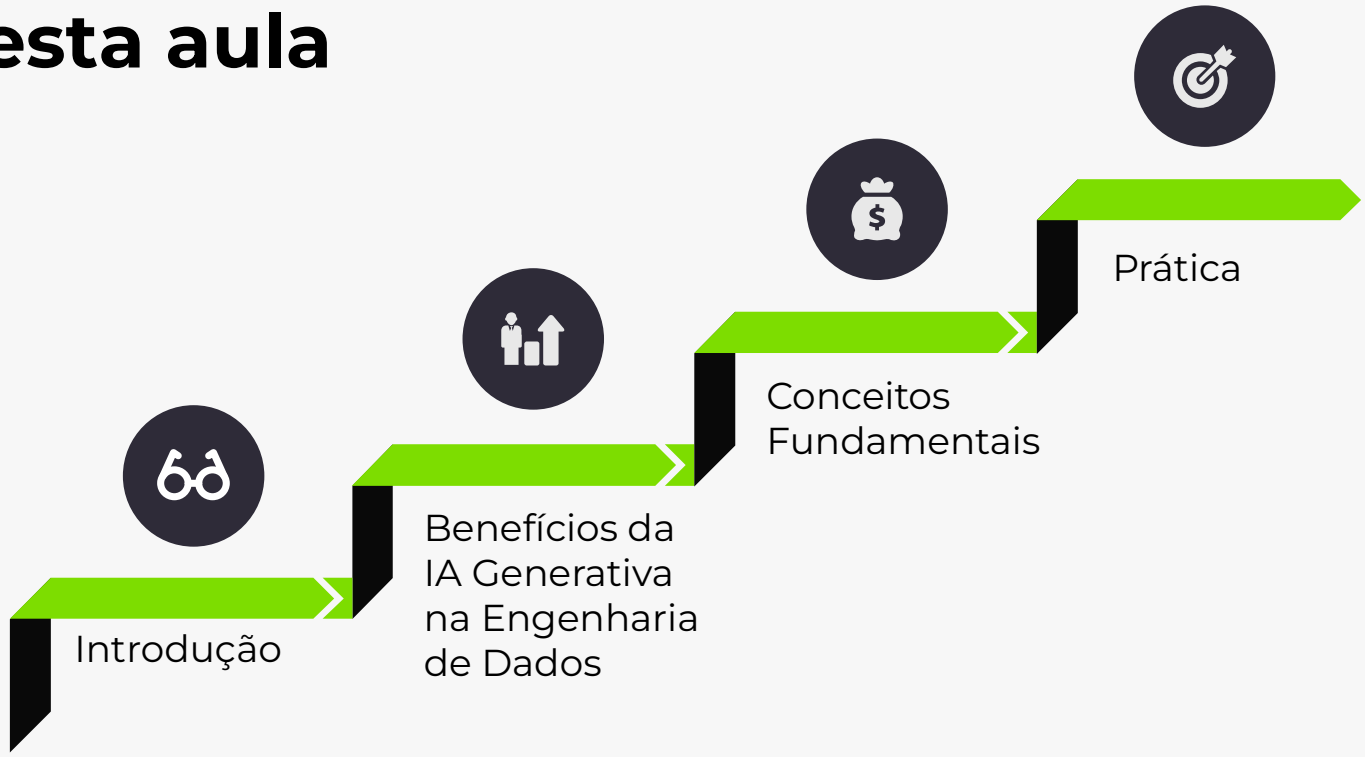
Módulo 3

Aula 01

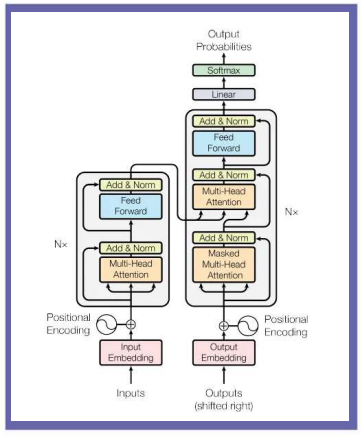
Prof. João Paulo Faria

https://www.linkedin.com/in/jpbfaria/

Nesta aula


Introdução
Benefícios da IA Generativa na Engenharia de Dados
Conceitos Fundamentais
Prática

Introdução



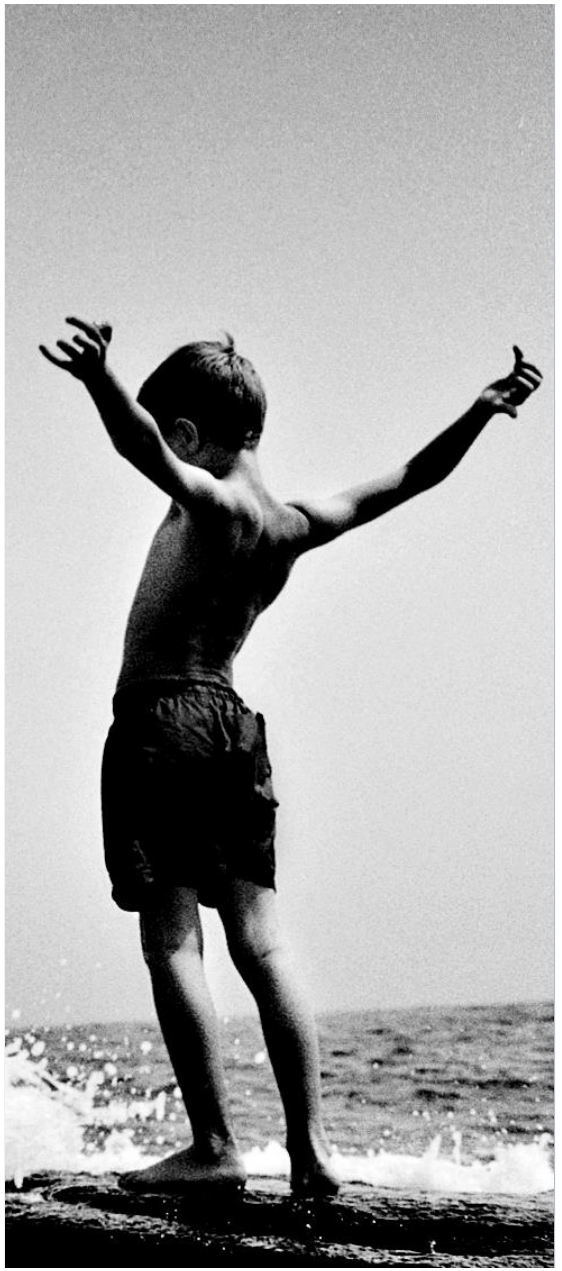
Transformer model extract

OpenAI

A OpenAI foi co-fundada em dezembro de 2015 por Elon Musk, Sam Altman, Greg Brockman, Ilya Sutskever, Wojciech Zaremba, entre outros.

Attention is all you need:

https://arxiv.org/pdf/1706.03762



Engenheiro de Dados IA Expert

Introdução

Google

Engenheiro de Dados IA Expert

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

O que é IA Generativa

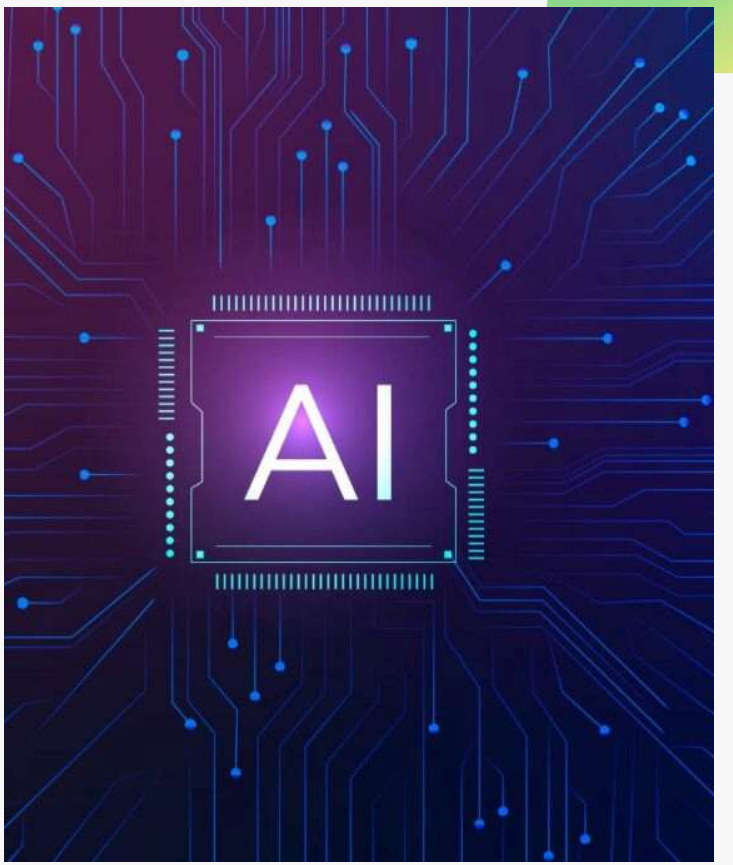
- Refere-se a uma categoria de modelos e ferramentas de IA projetadas para criar novos conteúdos, como texto, imagens, vídeos, música ou código.

- Usa uma variedade de técnicas – incluindo redes neurais e algoritmos de aprendizado profundo (Deep Learning) – para identificar padrões e gerar novos resultados.

O que a IAGen pode fazer?

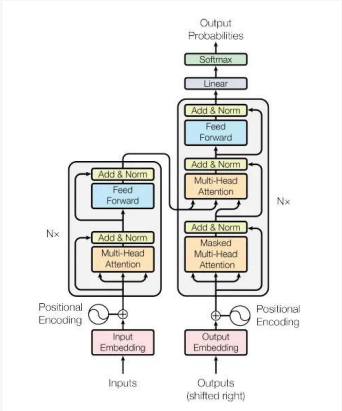
- Geração de Texto.
- Geração de Imagem.
- Geração de Vídeo.
- Geração de Código de Programação.
- Geração de Dados.
- Tradução de Idiomas.

Tudo em tempo real (a depender do poder de processamento)



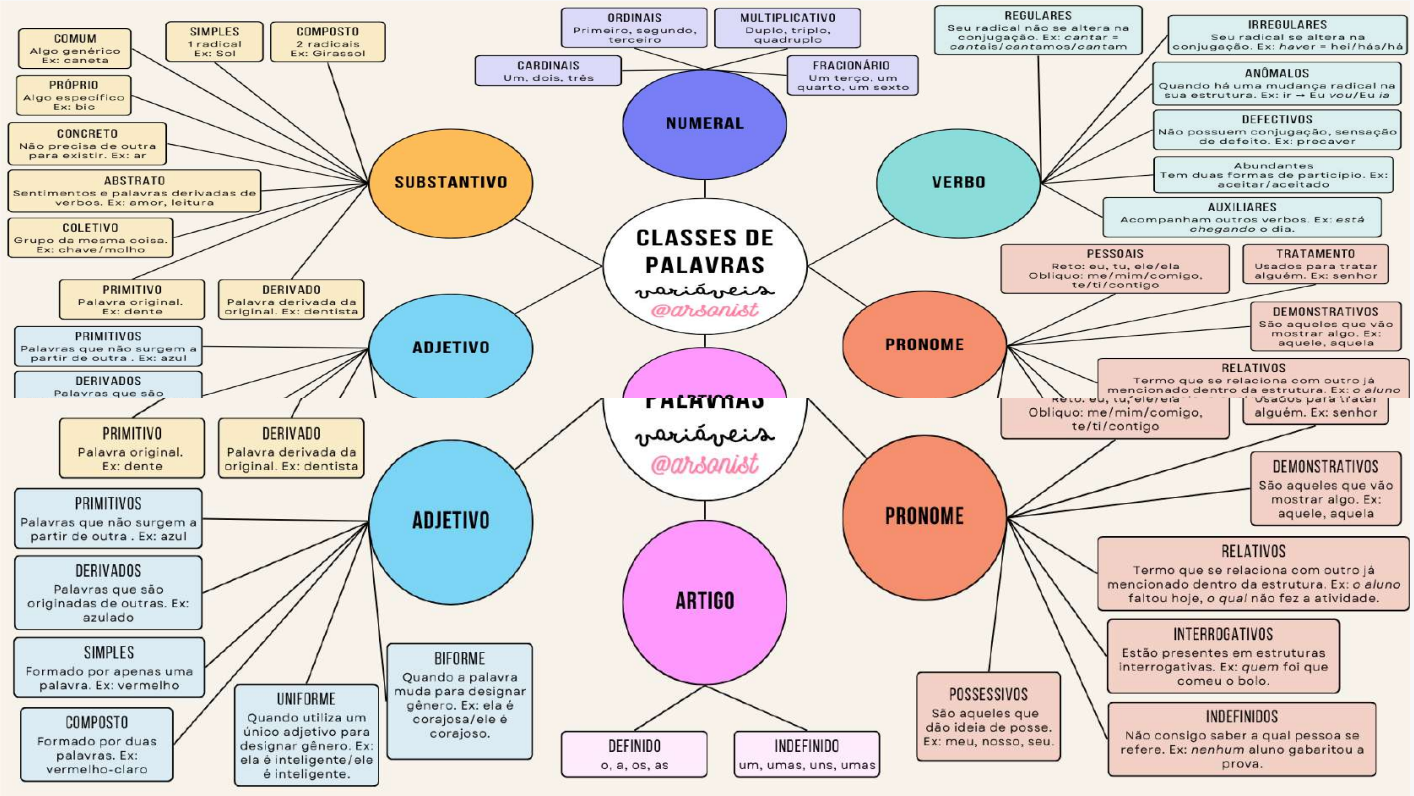
O que está por traz da IAG?

- Um modelo de linguagem grande (LLM). Um tipo de programa de inteligência artificial (IA) que pode reconhecer e gerar texto, entre outras tarefas.
- LLMs são treinados em grandes conjuntos de dados — daí o nome "grande".
- LLMs são construídos com algorítimos de ML: especificamente, um tipo de rede neural chamada modelo **transformador**.
- LLM é uma evolução do RNN.



Engenheiro de Dados IA Expert

IAGen gera uma palavra de cada vez



Portanto, é necessária uma classificação de cada palavra (etiquetamento)

Engenheiro de Dados IA Expert

IAGen gera uma palavra de cada vez

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Para cada palavra são aplicados milhões de parâmetros

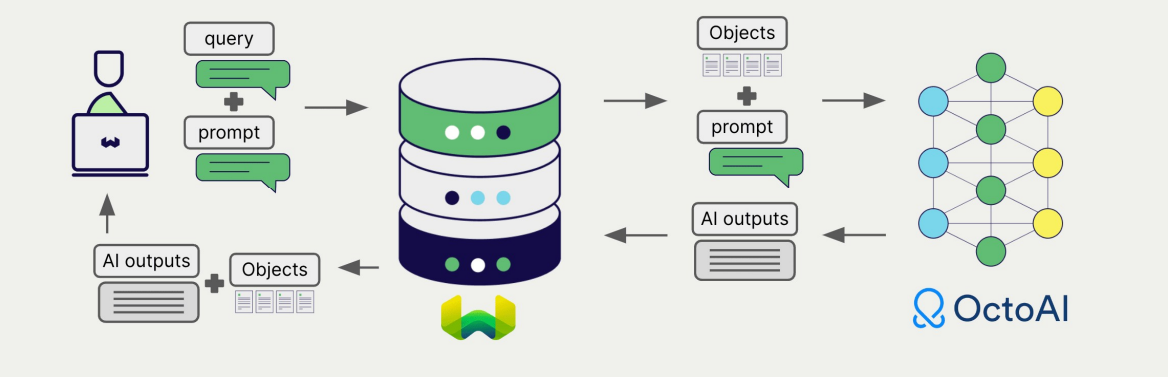
Faculdade XPe

O INPUT (prompt) se retroalimenta:

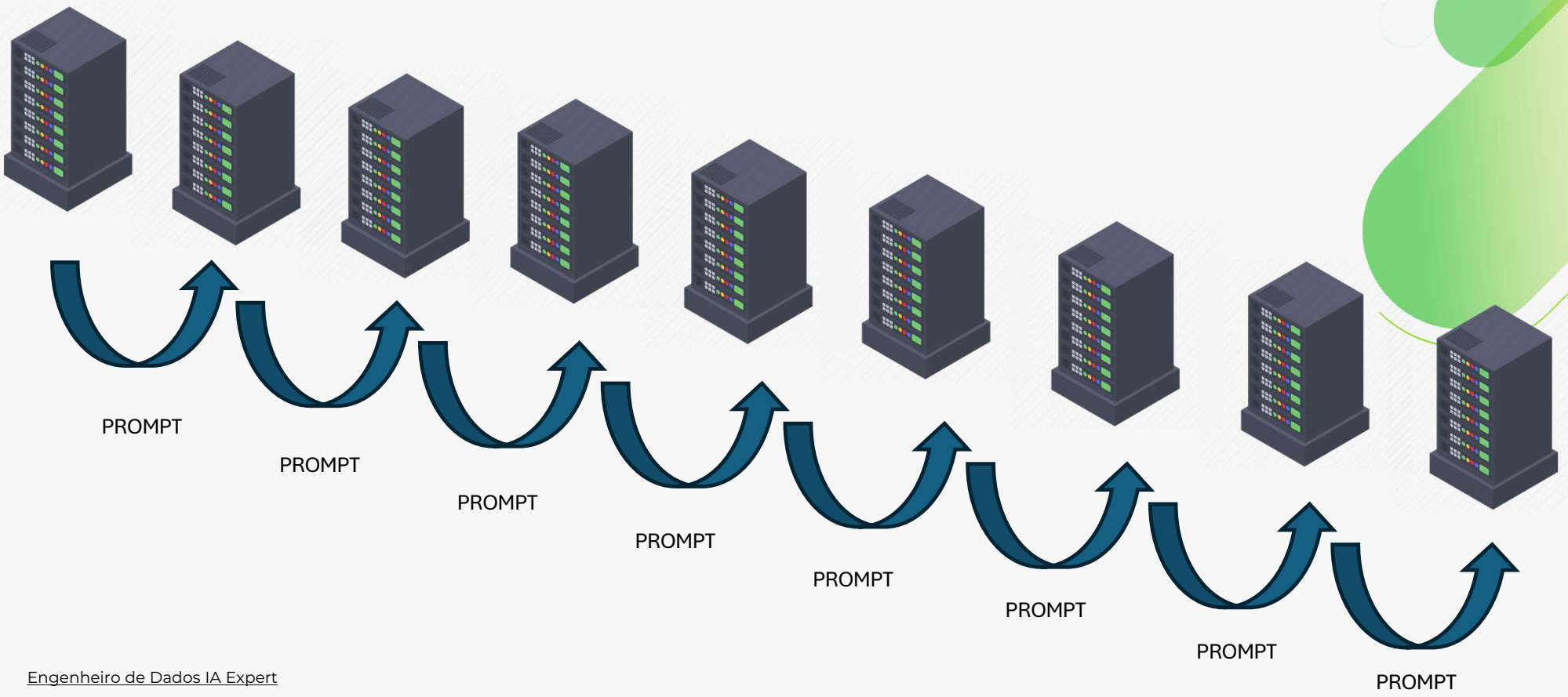
Prompt:

O céu é azul...

e	42%
claro	38%
escuro	10%
com	2%
acinzentado	1%
.	0.5%



E como tudo é tão rápido?

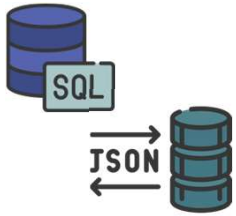


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

1 - Aceleração de Desenvolvimento

- **Geração de código para ETL/ELT (SQL, PySpark, dbt, Airflow, Spark Structured Streaming).**
- **Criação rápida de pipelines, DAGs, jobs de ingestão.**
- **Conversão de lógica de negócios em queries otimizadas, com redução de 30–60% no tempo de criação inicial de jobs ou modelos de dados.**

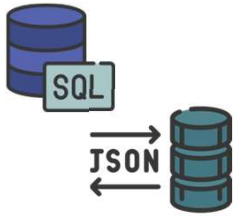


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

2 - Padronização e Melhores Práticas

- **Modelos podem sugerir padrões de nomenclatura, particionamento, incrementalidade, tipos, etc.**
- **Resultado: menor dívida técnica e menos divergência arquitetural entre times.**

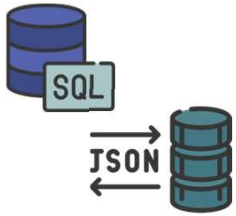


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

3 - Documentação Viva e (Data Discovery)

- **Geração automática de descrições de tabelas, colunas, lineage textual.**
- **Q&A semântico sobre o lake/warehouse - ex: “Quais tabelas têm métricas de churn?”.**
- **Impacto: onboarding mais rápido; menos dependência de “pessoas-oráculo”.**

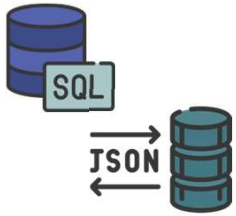


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

4 - Geração de Dados Sintéticos

- **Criação de datasets realistas para testes, POCs e cenários de privacidade. (vamos fazer hoje!)**
- **Balanceamento de classes para modelos downstream.**
- **Cuidados: evitar vazamento de dados sensíveis; usar técnicas de privacidade (DP, k-anonymity simulada).**

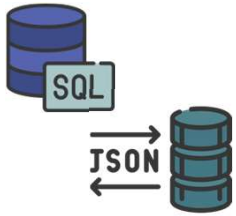


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

5 - Testes e Qualidade de Dados

- **Geração de testes (ex: Great Expectations, Soda) a partir de perfis de dados.**
- **Sugestão de constraints (unicidade, ranges, foreign keys lógicas).**
- **Geração de queries de validação para regressões após mudanças de schema.**
- **Impacto: aumento de cobertura de testes com menor esforço manual.**

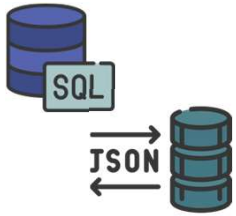


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

6 - Observabilidade e Anomalias

- **Explicação em linguagem natural de quedas de volume, spikes de latência, drift em métricas.**
- **Criação de códigos de mitigação automaticamente.**
- **Ex: “Volume do tópico Vendas caiu 47% vs média dos últimos 7 dias; possível causa: falha no job upstream X às 02:15.”**

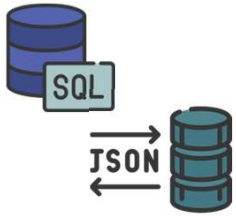


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

7 - Otimização de Performance e Custos

- **Reescrita de queries SQL ou Spark para reduzir shuffle, scans completos, explosões de join.**
- **Sugestão de particionamento, clustering, Z-order, materialized views.**
- **Identificação de tabelas “zumbi” e jobs redundantes.**
- **Impacto típico: economias de 10–40% em custos de warehouse/lake compute.**

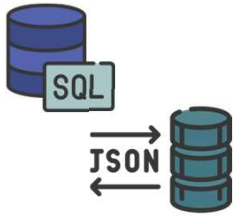


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

8 - Governança, Compliance e Segurança

- **Classificação semiautomática de PII/PHI a partir de padrões e contexto semântico.**
- **Geração de políticas (ex: Lakehouse row-level, column masking) em formato de código.**
- **Resumos de impacto de uma mudança de schema sobre domínios regulados.**
- **Cuidados: revisão humana obrigatória para classificações sensíveis.**

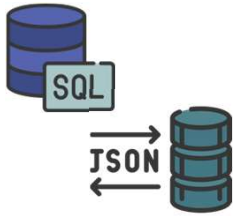


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

9 - Automação de Mapeamentos/Integrações

- **Mapeamento semântico entre fontes heterogêneas (ERP A vs ERP B).**
- **Geração de transformações para normalizar taxonomias (ex: códigos de países, unidades).**

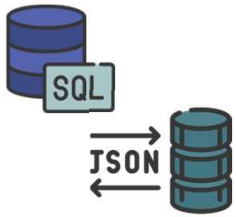


Faculdade XPe

Benefícios do uso da IAGen na Engenharia de Dados

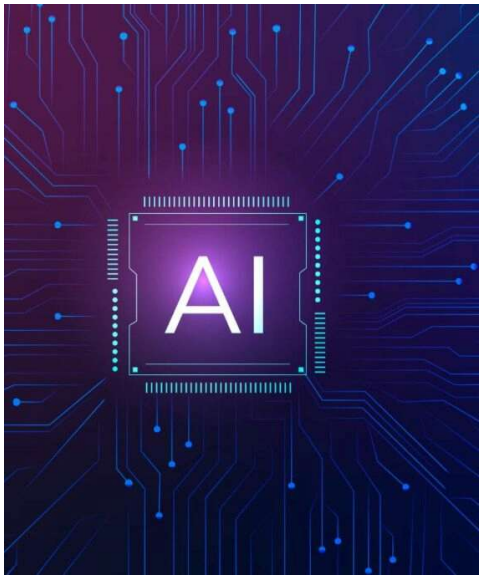
10 - Melhoria de Colaboração e Transferência de Conhecimento

- **Assistentes conversacionais integrados a repositórios de dados e código.**
- **Conversões de reuniões técnicas gravadas em decisões acionáveis + diffs sugeridos.**



Faculdade XPe

Outros:



Imagine os seguintes cenários:

- **Protótipos e Ideação**
- **Suporte Multilíngue e Democratização**
- **Redução de Erros Humanos**
- **Curadoria de Pipelines Existentes**
- **Planejamento de Capacity e SLAs**
- **Aceleração de Migrações Tecnológicas**

Faculdade XPe

Métricas de Valor

- Time-to-first-production-pipeline: -40%.
- Percentual de queries otimizadas automaticamente: >25%.
- Redução de incidentes de qualidade críticos: -30%.
- Aumento de cobertura de testes de dados: +50%.
- Custo compute por TB transformado: -15–35%.

Faculdade XPe

Conceitos Importantes

1 Preprocessamento de Dados:

> É o conjunto de etapas sistemáticas aplicadas aos dados brutos logo após (ou junto com) a ingestão, para torná-los consistentes, confiáveis, estruturados e prontos para uso em camadas posteriores. É a “ponte” entre dado cru e dado utilizável.

Faculdade XPe

Conceitos Importantes

1 Preprocessamento de Dados:

> Reduzir impurezas: remover ruído, linhas inválidas, duplicidades.

Tornar os dados interoperáveis: tipos corretos, formatos de data, codificações, normalização de unidades.

Preservar rastreabilidade: manter lineage e histórico de alterações.

Faculdade XPe

Conceitos Importantes

2 Redução da Dimensionalidade:

> O Processo de diminuir o número de variáveis (features) usadas para representar os dados, preservando (tanto quanto possível) a informação relevante (estrutura, variância, separabilidade, poder preditivo).

Faculdade XPe

Conceitos Importantes

2 Redução da Dimensionalidade:

> É uma resposta prática ao “mal da dimensionalidade” (curse of dimensionality), onde dados com muitas colunas se tornam esparsos, modelos ficam mais lentos, overfitting aumenta e a interpretação piora.

Faculdade XPe

Conceitos Importantes

3 Normalização:

> Normalização pode significar coisas diferentes em Engenharia de Dados / Dados & ML. O termo é ambíguo, então demonstro os principais contextos.

Faculdade XPe

Normalização

De Banco de Dados Relacional

- Reduzir redundância e evitar anomalias de inserção, atualização e exclusão em modelos transacionais. Como: Quebrar uma tabela “larga” em múltiplas tabelas relacionadas aplicando regras (Formas Normais).

Faculdade XPe

Normalização

Padronização de Dados (Data Standardization)

- Ter valores com formatos, unidades e codificações consistentes. Exemplos:
  Datas: converter tudo para ISO 8601 UTC.
  Países: usar ISO 3166 (BR, US) ao invés de “Brasil”, “BRA”, “Brazil”.

Faculdade XPe

Normalização

Normalização de Features (Feature Scaling)

- Ajustar escala dos atributos numéricos para que algoritmos sensíveis a magnitude ou distância funcionem melhor.

Faculdade XPe

Normalização

Normalização de Texto (Text Normalization / NLP)

Reduzir variação superficial para facilitar análise.

Etapas comuns:

- Lowercase.
- Remover pontuação / símbolos.
- Normalizar acentos (NFKD).
- Lematização ou stemming.
- Remoção de stopwords.

Faculdade XPe

Normalização

Normalização Estatística

- Escalar séries temporais diferentes para comparar padrões (z-score por série).

- Normalizar “participação” para proporções (dividir por total).

- Softmax (transforma logit em distribuição de probabilidade).

- Normalização em Vetores (Search / Similaridade)

Demonstração Prática!

Contatos

E-mail: joao.faria.@xpe.edu.br

LinkedIn: https://www.linkedin.com/in/jpbfaria