

SQL WEEK RECIFE

Conference - 6^a edição

05/02 até 09/02





APOIO

- LS Treinamentos
- Consulta BD
- Lucrécio Paes (CO-HOST)



QUEM EU?



- Carreira iniciada em 2005 na área de TI/Dados.
- Instrutor de Treinamentos de Banco de Dados e T-SQL
- MCT Microsoft 2023 – 2024
- Azure Data Engineer na 4Zoom Alocado na Nestlé
- Algumas Certificações na carreira (Microsoft, Python, Databricks, Itil e Cobit)
- Especialista em Modelagem de Dados e Arquitetura de Dados
- Vasta experiência em Tecnologias de Banco de Dados SQL Server, SSIS, SSAS, Azure, Azure Data Factory, Databricks, Python e Spark/PySpark
- Possuo blog, Instagram, Twitter Dba Assists...sigam lá!



Gabriel Quintella





APLICANDO DATA QUALITY EM UM PIPELINE DE DADOS

Data Quality, o que é?



Termo utilizado para definir...

Nível de **qualidade de dados** dentro de uma companhia!

Data Quality

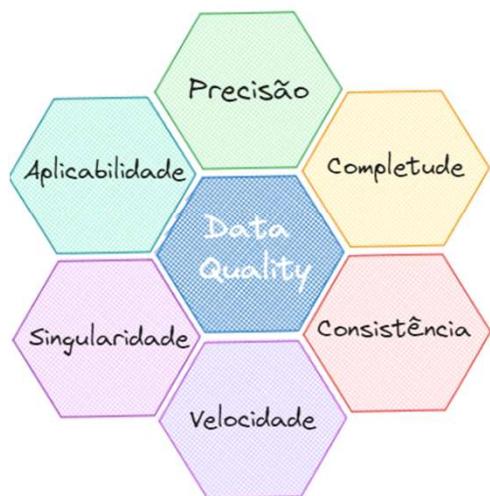
Veracidade

Confiabilidade dos dados. Devem ter qualidade, consistência, origens conhecidas e não inventados (opinião)...

Diz respeito às **informações** armazenadas...

Podem ser **qualitativas ou quantitativas**, mas o que realmente define se são boas, ou não, é sua relevância para solucionar o problema que as fizeram ser coletadas!

Pilares do Data Quality



Precisão - Toda informação deve descrever ou refletir o mundo real, ou seja, devem ser fiéis à realidade;

Completude - Os dados não podem estar fragmentados;

Consistência - Dado consistente é aquele que, quando comparado com outras amostras, se confirma e não flutua. Seus dados não pode dizer que para um mês o valor é de R\$20 mil e na outra semana apontam que esse valor é de R\$5 mil. A consistência é essencial para garantir informação de qualidade.

Velocidade - Com que velocidade você tem acesso aos seus dados? Dentro de uma governança de dados (Data Governance) efetiva, a agilidade na entrega das informações é fundamental para a operação. A qualidade do dado também envolve a velocidade necessária de acesso aos diferentes níveis e hierarquias responsáveis por utilizá-lo.

Singularidade - Informações duplicadas podem resultar em erros críticos, pois podem não refletir a realidade.

Aplicabilidade - O dado que você armazena tem que ser utilizável pois caso contrário é desperdício de dinheiro e tempo.

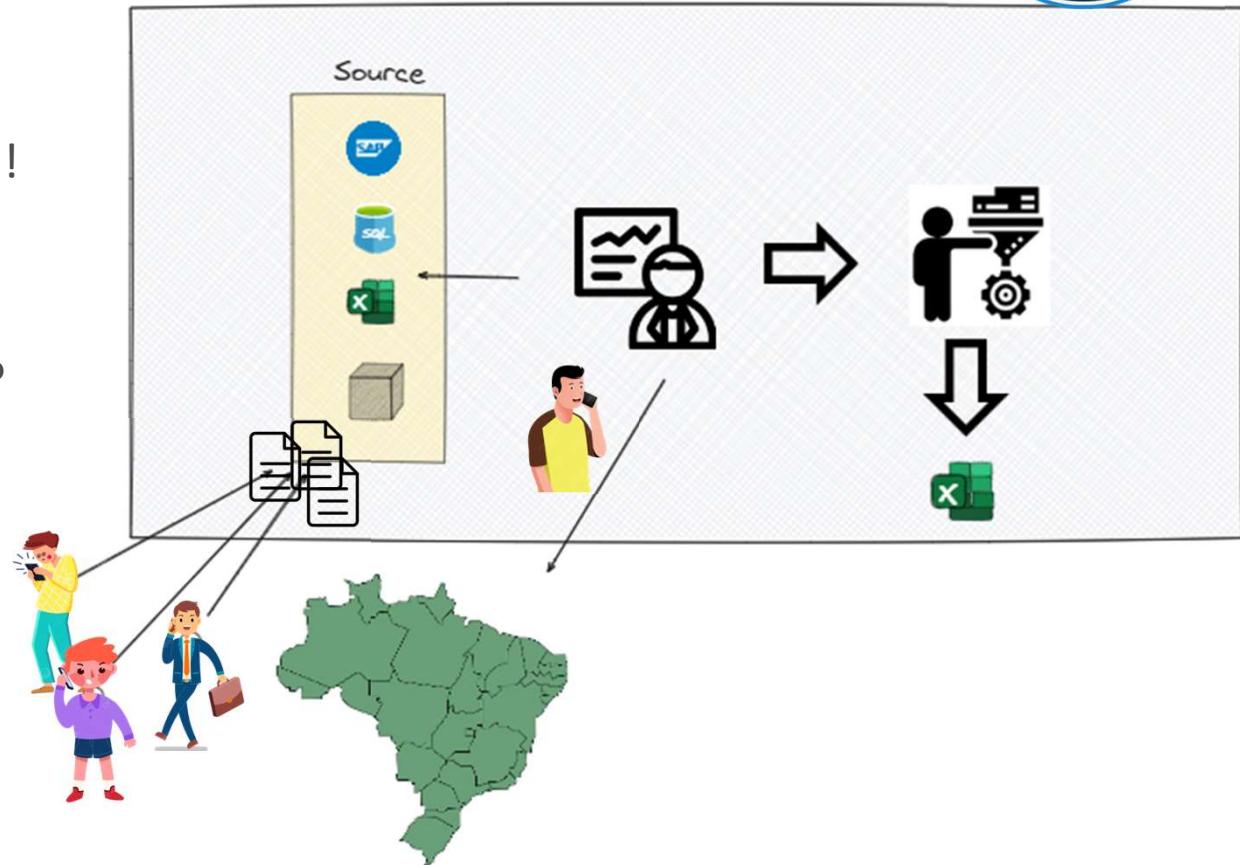
Fonte: <https://www.sysvision.com.br/blog/data-quality-o-que-e-qualidade-de-dados>

Como que era o cenário?



Missão: Automatizar a geração dos indicadores operacionais da empresa!

Como que esses indicadores são gerados?

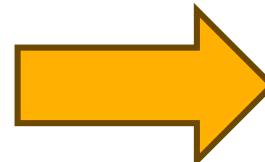


Nada muito assustador...



Esses indicadores eram compostos por diversas fontes de dados:

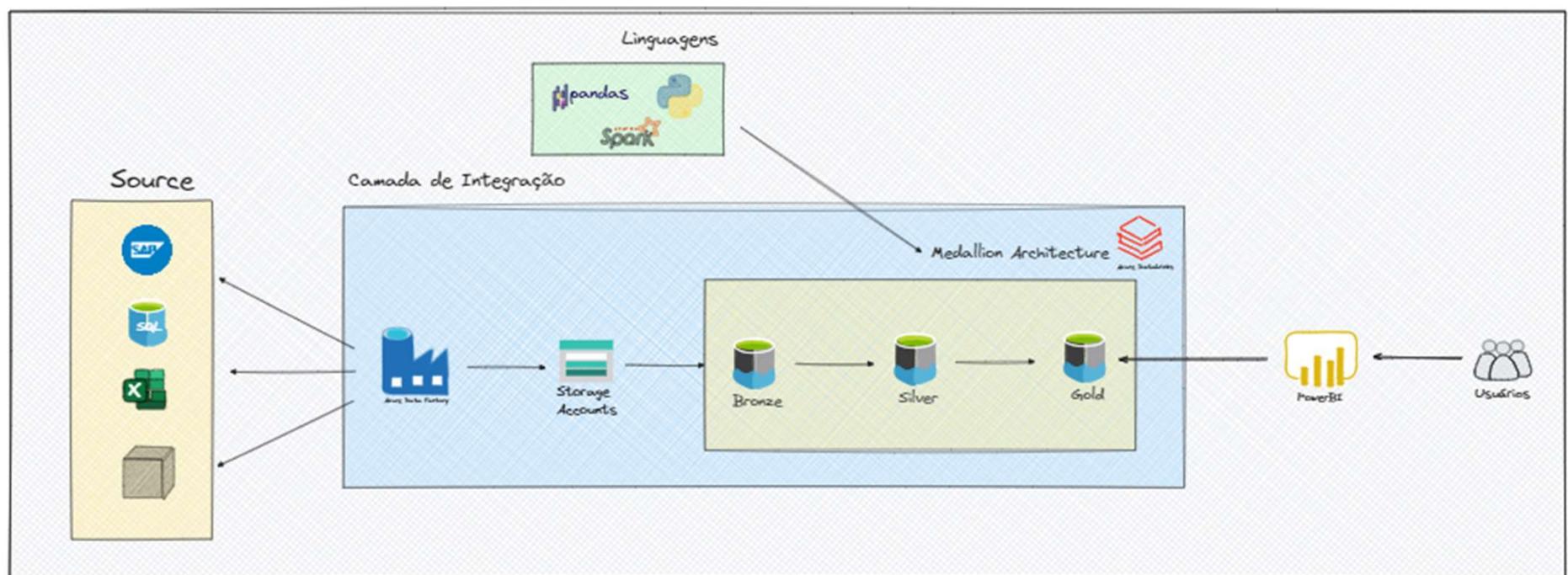
- 1 – Arquivos CSV e XLSX
- 2 – Cubo
- 3 – SQL Server
- 4 – Google Sheets
- 5 – ERP SAP



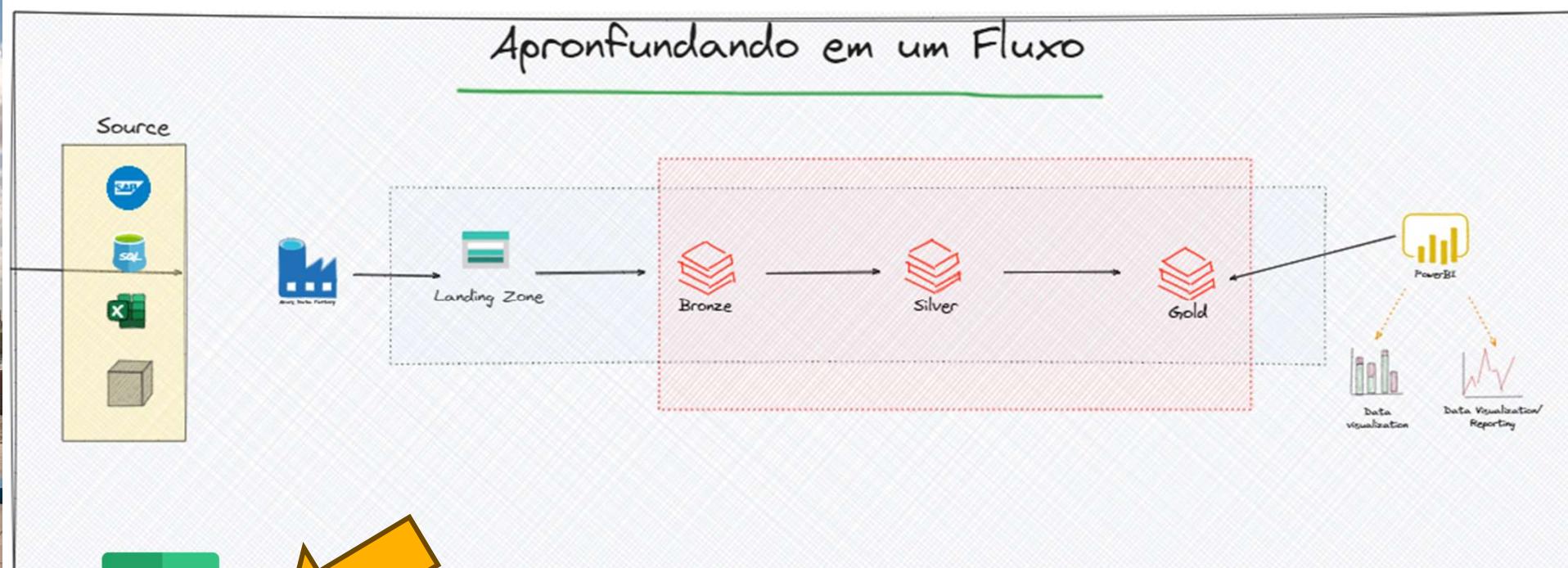
Gerava um Excel maroto com 89 abas...



Arquitetura Proposta

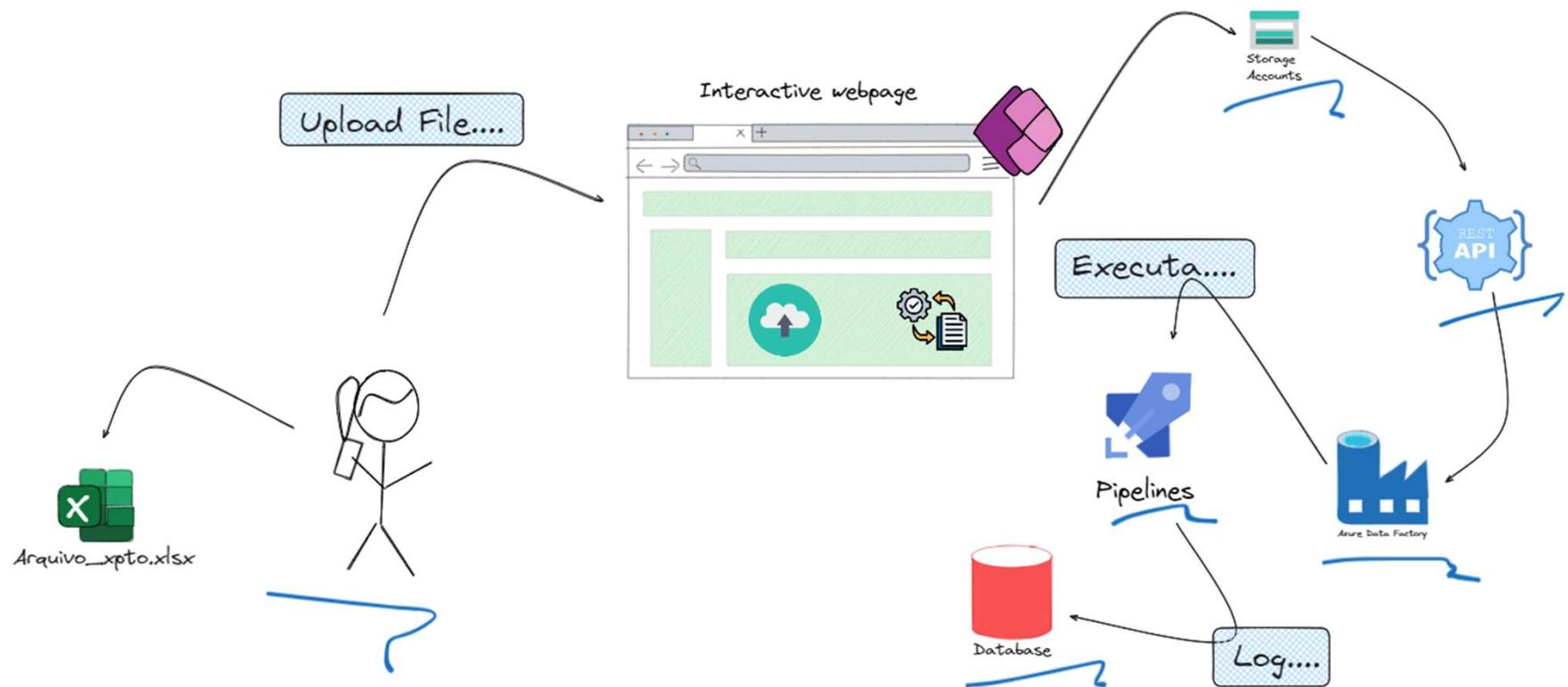


Conhecendo um Fluxo...



Arquivo_xpto.xlsx

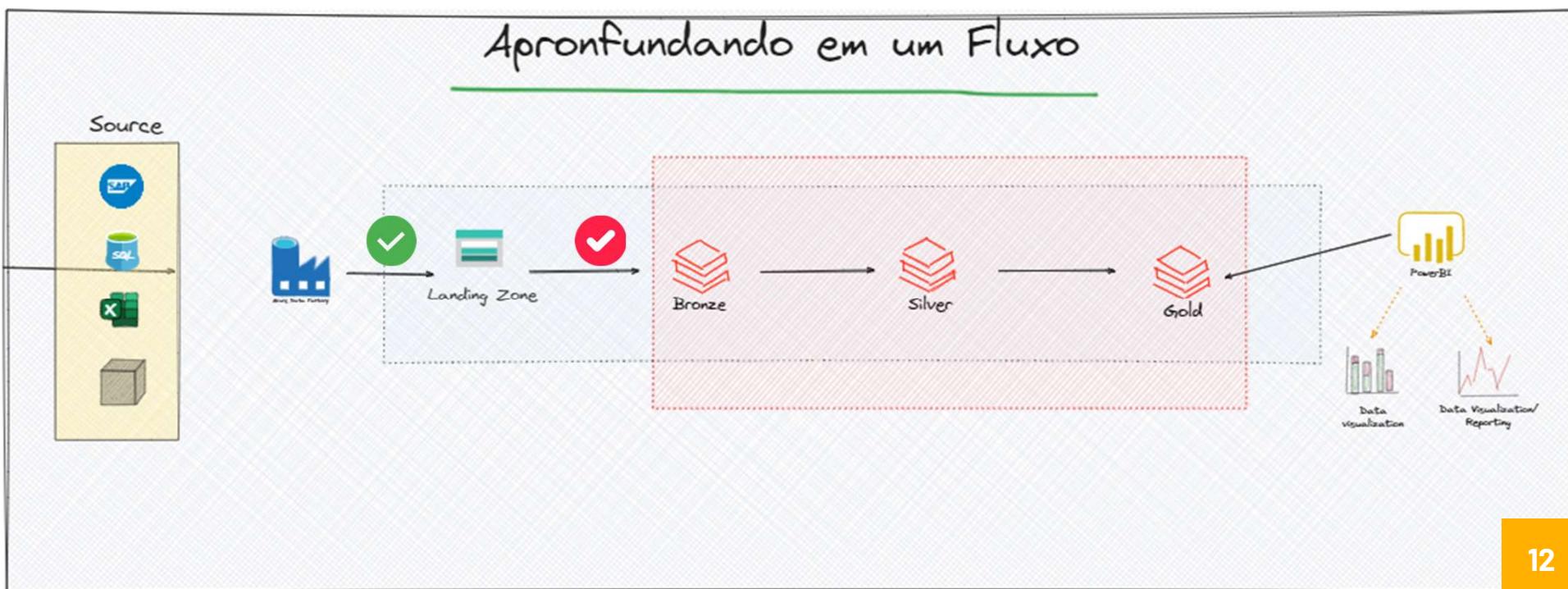
Processando...



Pammmm!!!



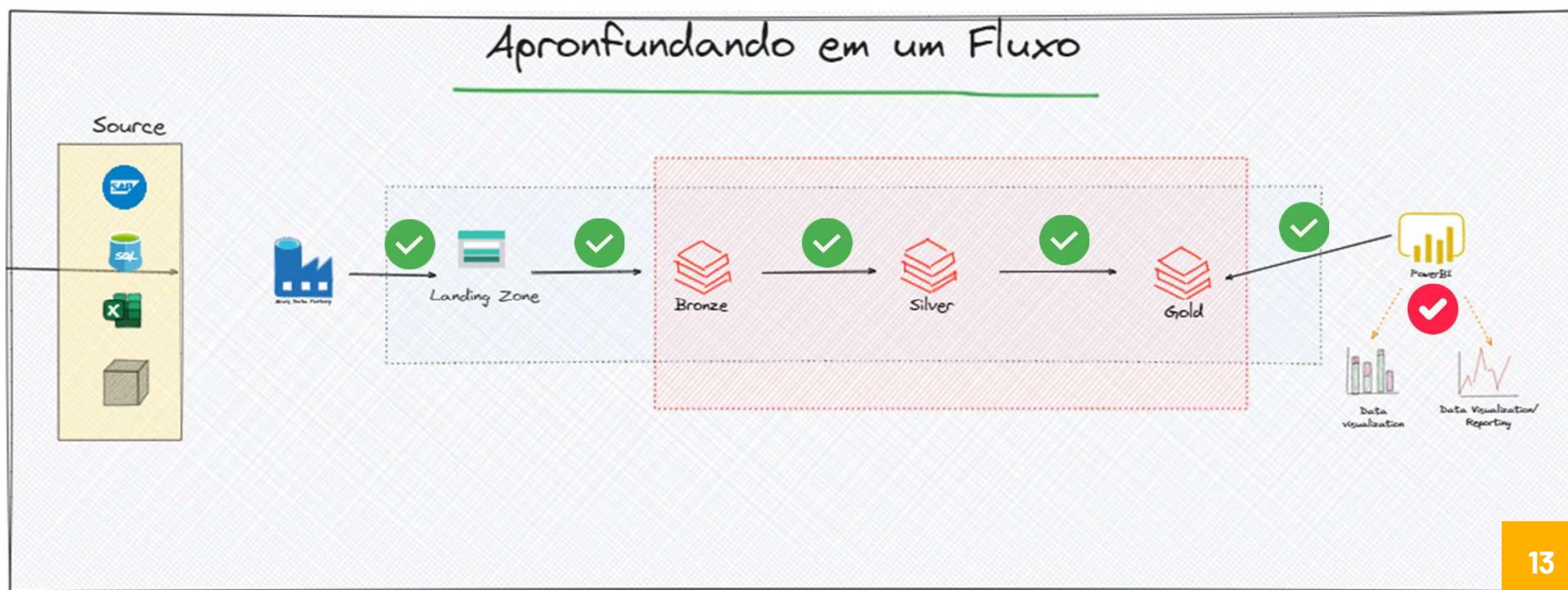
Apronfundando em um Fluxo



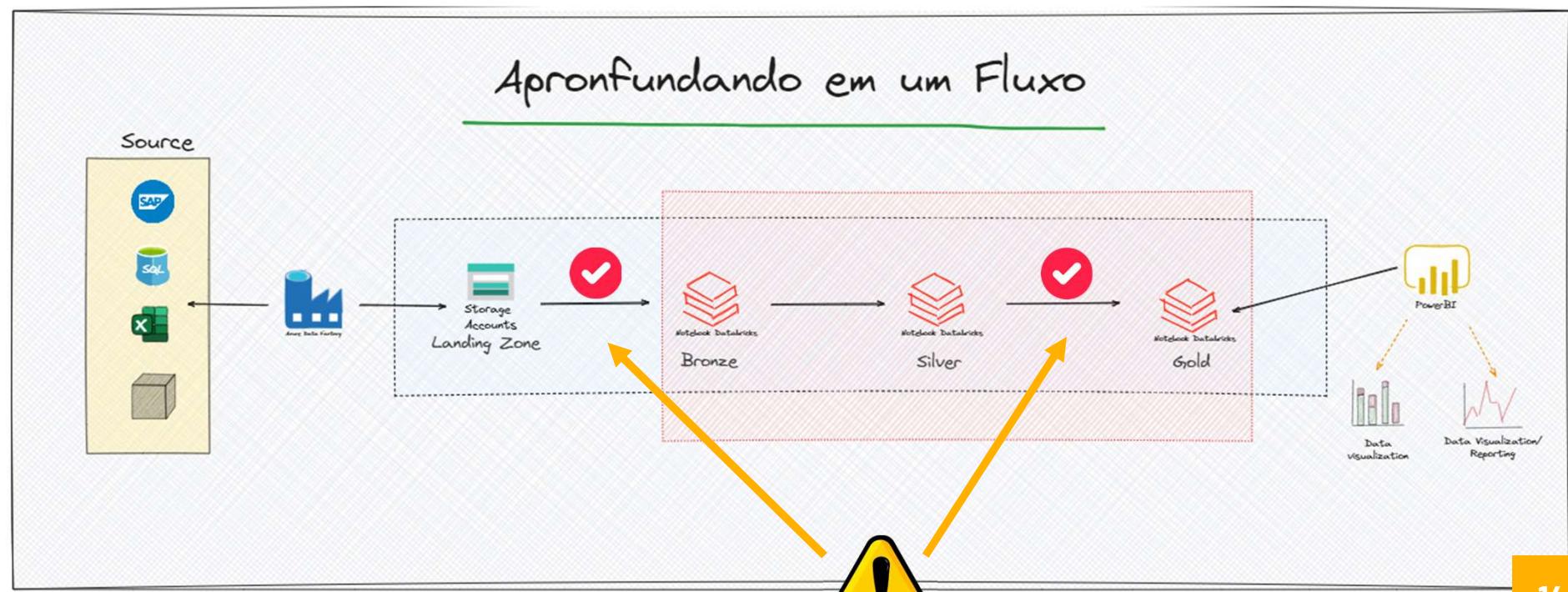
Pammmm!!!



Apronfundando em um Fluxo



Pontos Problemáticos



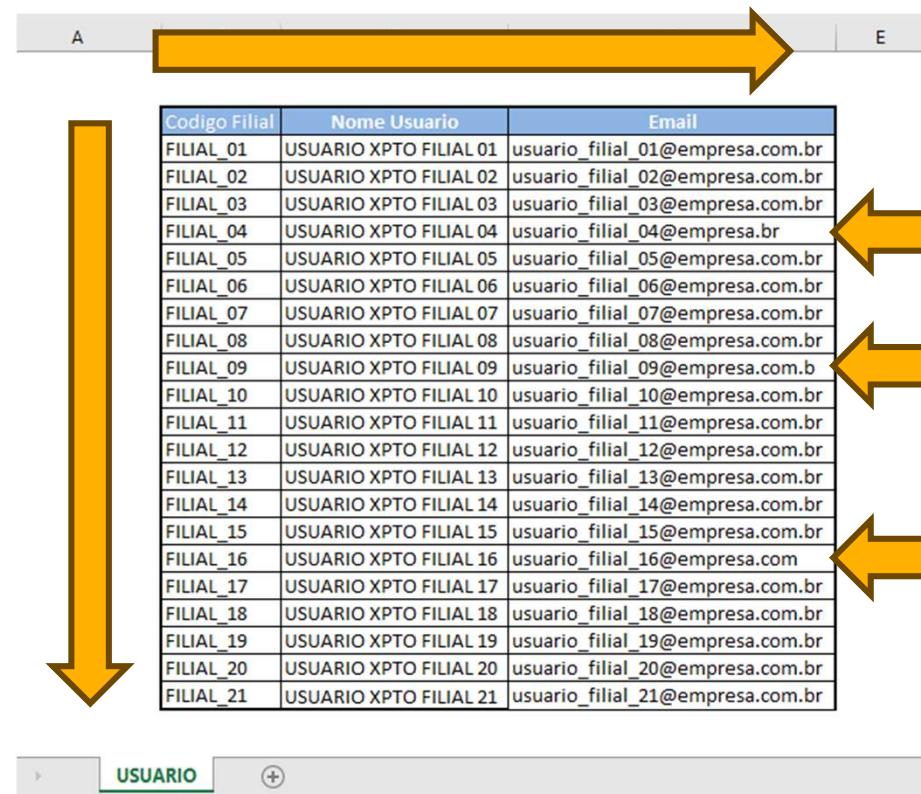
Pontos Problemáticos



The screenshot shows an Excel spreadsheet with data for 21 branches (Filial XPTO 01 to Filial XPTO 21). The columns represent months from Janeiro to Dezembro, plus a YTD column. A large yellow arrow points from the bottom left towards the table, and another large yellow arrow points from the bottom right towards the table. A cartoon detective character is positioned to the right of the table, holding a magnifying glass.

Código Filial	Filial	Janeiro	Fevereiro	Março	Abril	Mai	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro	YTD
FILIAL_01	FILIAL XPTO 01	27,929	26,929	25,929	25,929	25,929	25,929	27,929	28,929	27,929	28,929	28,929	27,929	329,147
FILIAL_02	FILIAL XPTO 02	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000	4,000	48,000
FILIAL_03	FILIAL XPTO 03	19,580	19,580	20,580	20,580	20,580	20,580	22,580	21,580	21,580	23,580	21,580	21,580	253,900
FILIAL_04	FILIAL XPTO 04	8,713	8,713	8,713	8,713	8,713	8,713	8,713	8,713	8,713	8,713	7,713	7,713	102,100
FILIAL_05	FILIAL XPTO 05	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	2,000	24,000
FILIAL_06	FILIAL XPTO 06	21,069	21,069	21,069	21,069	21,069	21,069	21,069	20,069	20,069	21,000	21,000	21,000	25,000
FILIAL_07	FILIAL XPTO 07	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	5,000	50,000
FILIAL_08	FILIAL XPTO 08	17,242	17,242	17,242	16,242	16,242	16,242	16,242	16,242	16,242	16,242	16,242	16,242	197,000
FILIAL_09	FILIAL XPTO 09	7,000	7,000	7,000	7,000	8,000	8,000	8,000	8,000	8,000	8,000	8,000	8,000	92,000
FILIAL_10	FILIAL XPTO 10	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	84,000
FILIAL_11	FILIAL XPTO 11	4,000	4,000	4,000	4,000	4,000	4,000	4,000	5,000	5,000	5,000	5,000	5,000	57,000
FILIAL_12	FILIAL XPTO 12	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	7,000	6,000	7,000	70,000
FILIAL_13	FILIAL XPTO 13	19,322	19,322	19,322	19,322	19,322	18,322	19,322	18,322	19,422	19,322	19,322	19,322	19,322
FILIAL_14	FILIAL XPTO 14	14,500	14,500	14,500	14,500	14,500	14,500	14,500	14,500	14,500	14,500	14,500	14,500	15,500
FILIAL_15	FILIAL XPTO 15	13,000	12,000	12,000	12,000	12,000	12,000	12,000	11,000	12,000	12,000	12,000	12,000	12,000
FILIAL_16	FILIAL XPTO 16	6,000	6,000	6,000	6,000	6,000	7,000	7,000	7,000	9,000	8,000	8,000	8,000	8,000
FILIAL_17	FILIAL XPTO 17	9,000	9,000	9,000	9,000	9,000	9,000	9,000	9,000	9,000	9,000	9,000	9,000	108,000
FILIAL_18	FILIAL XPTO 18	9,769	9,769	9,769	9,769	9,769	9,769	9,769	9,769	9,769	10,769	10,769	10,769	120,225
FILIAL_19	FILIAL XPTO 19	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	120,000
FILIAL_20	FILIAL XPTO 20	3,000	3,000	3,000	3,000	3,000	3,000	4,000	4,000	4,000	4,000	4,000	4,000	42,000
FILIAL_21	FILIAL XPTO 21	215,123	213,123	213,123	212,123	213,123	213,123	219,123	217,123	220,223	224,054	220,054	221,054	2601,367

Pontos Problemáticos

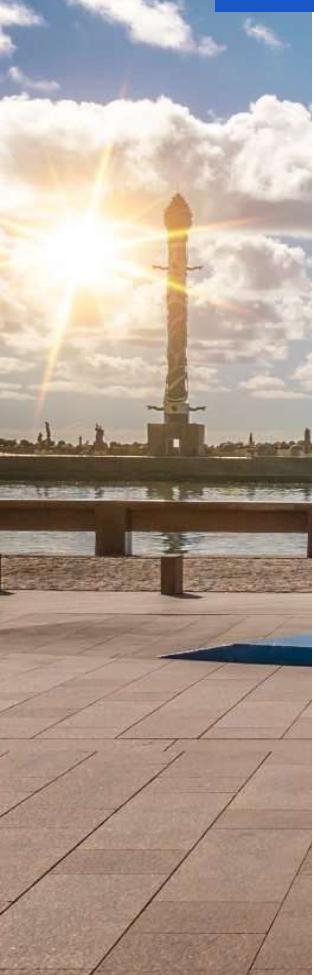


Medição



Principais pontos de ruptura do processo:

1. Alteração de estrutura (50%);
2. Dado com formato errado (20%);
3. Erro de preenchimento (15%);
4. Ausência de Dados (10%);
5. Outros (5%)





Ajustes Necessários



Principal Erro – Não aprofundar no conhecimento das fontes e realizar a criação do processo AS-IS

Onde melhorar:

1. Criar uma rotina eficiente de Data Quality;
2. Definir uma estrutura fixa para as fontes (arquivos);
3. Validação dos tipos de dados, em caso de inconsistência rejeita o arquivo;
4. Validação de variação aceitável.

Como fazer isso funcionar?



1 – Arquivo

- Todo arquivo tem um nome fixo;
- Validar existência do arquivo no diretório cadastrado;
- Caso formato IGUAL xlsx, validar se existe uma aba com o nome informado;
- Caso formato IGUAL csv, validar se o arquivo está delimitado conforme informado;

2 – Estrutura

- Validar a estrutura cadastrada para o arquivo;
- Validar o tipo de dado para cada coluna conforme cadastro;
- Validar colunas NULAS;
- Validar coluna CHAVE;

3 - Dado

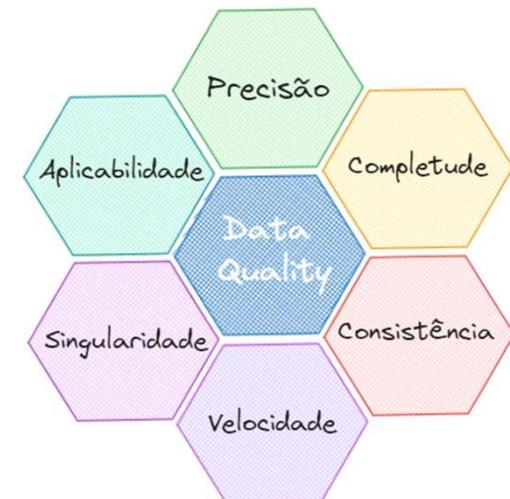
- Validar dados conforme coluna TIPO_DADO

4 – Variação de Valores

- Validar coeficiente de variação (Usuário pode baypassar essa etapa!)

5 – Notificar qualquer inconsistência

Pilares do Data Quality



Como fazer isso funcionar?



1 – Arquivo

Todo arquivo tem um nome fixo;
Validar existência do arquivo no diretório cadastrado;
Caso formato IGUAL [xlsx](#), validar se existe uma aba com o nome informado;
Caso formato IGUAL [csv](#), validar se o arquivo está delimitado conforme informado;

Arquivo	Diretório	Formato	Aba	Delimitador	Coeficiente_Variacao
XPTO_ARQUIVO	/asaempresaxpto/vendas/	XLSX	Plan1		0,1

Arquivo	E-mail Responsável
XPTO_ARQUIVO	fulano@empresa.com

5 – Notificar qualquer inconsistência

2 – Estrutura

Validar a estrutura cadastrada para o arquivo;
Validar o tipo de dado para cada coluna conforme cadastro;
Validar colunas NULAS;
Validar coluna CHAVE;

Seqüencia	Coluna	DataType	Nulo	Chave	Tipo_Dado
1	CODIGO_FILIAL	INT	0	1	INTEIRO
2	NOME_FILIAL	VARCHAR(100)	0	1	STRING
3	DATA_INICIAL_MES	DATETIME	0	1	DATETIME
4	DATA_FINAL_MES	DATETIME	0	1	DATETIME
5	VALOR_FATURAMENTO	DECIMAL(23,3)	0	0	DECIMAL
6	LATITUDE	DECIMAL(23,3)	1	0	DECIMAL
7	LOGITUDE	DECIMAL(23,3)	1	0	DECIMAL

4 – Variação de Valores

Validar coeficiente de variação (Usuário pode bypassar essa etapa!)

3 - Dado

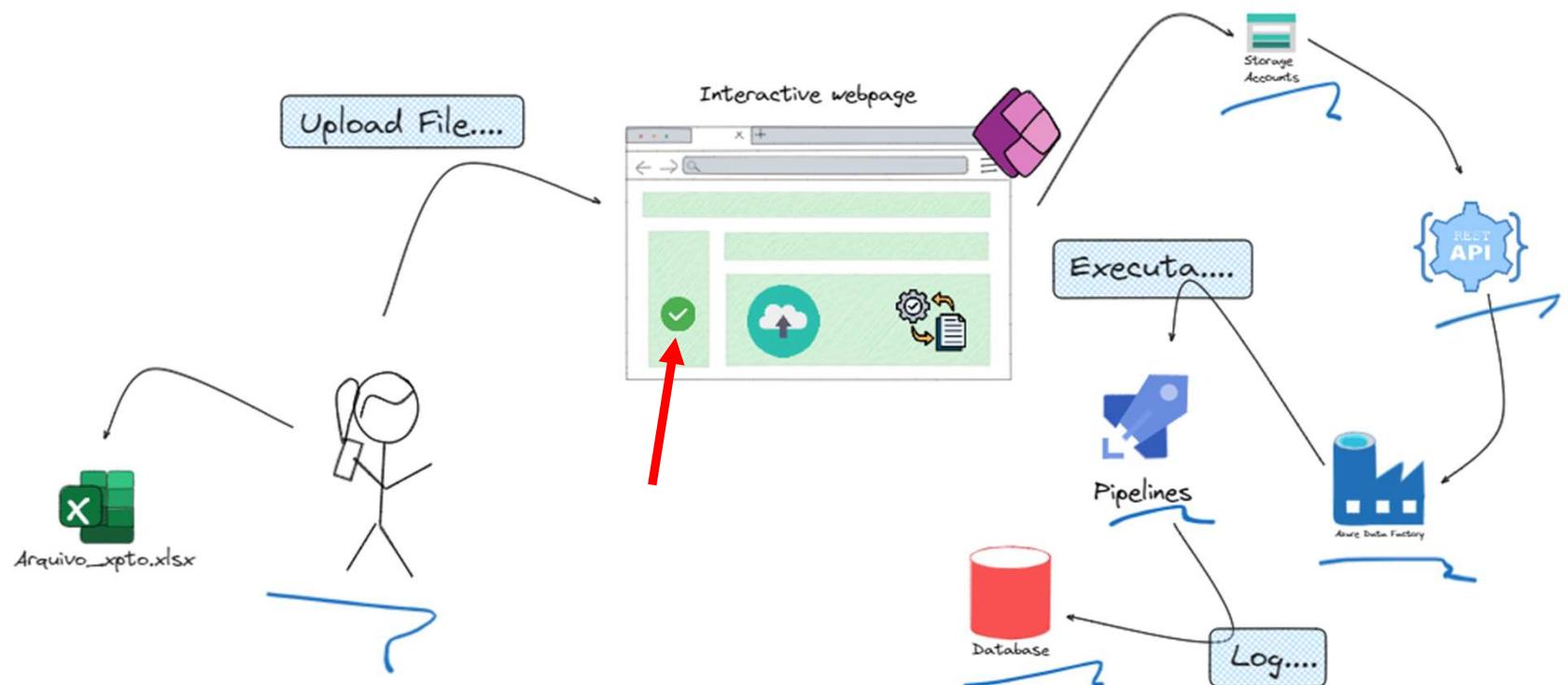
Validar dados conforme coluna TIPO_DADO

Nulo	
0	Não
1	Sim

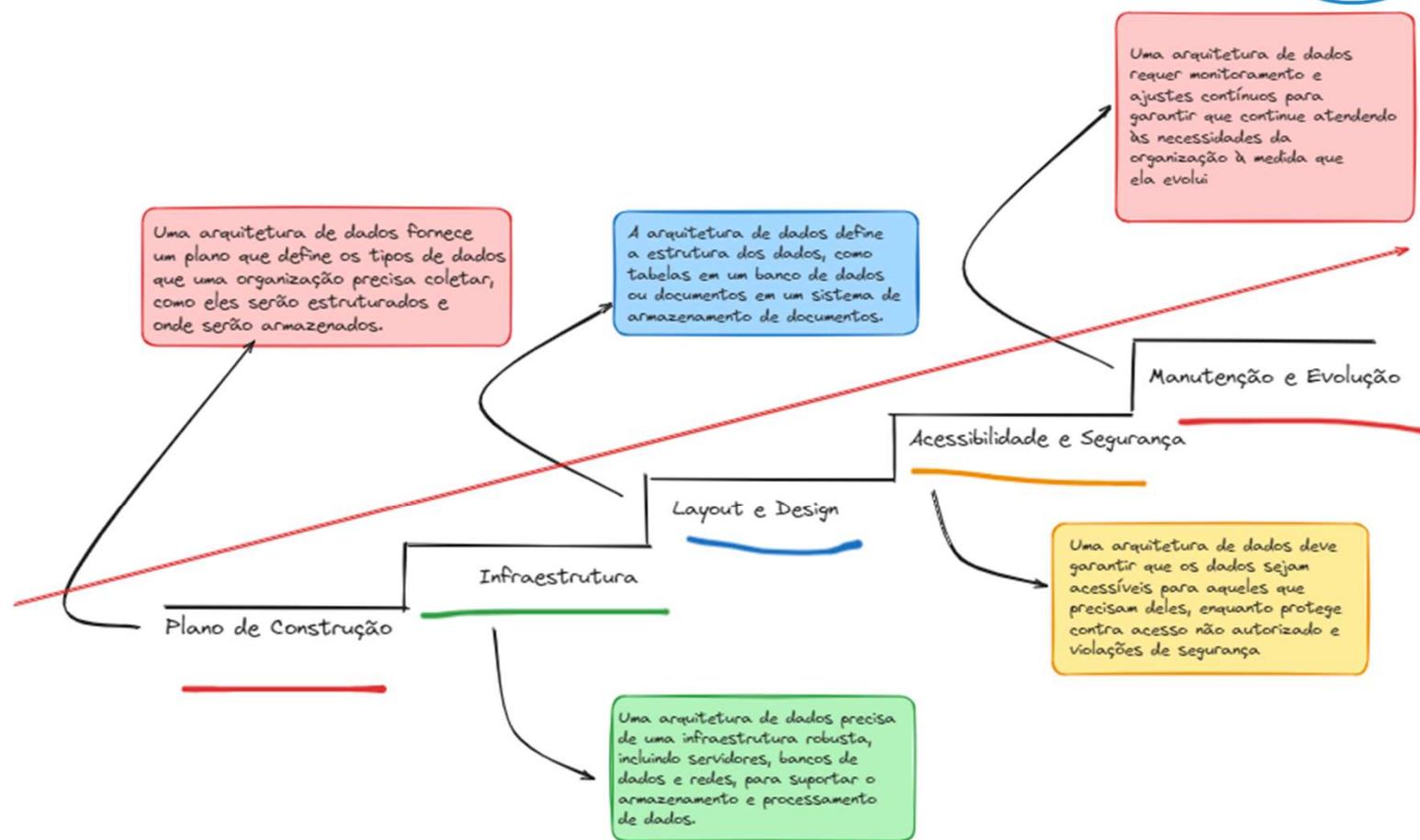
Chave	
0	Não
1	Sim

DataType	TIpo_Coluna
SMALLINT/INT/BIGINT	INTEIRO
VARCHAR/CHAR	STRING
DATETIME	DATETIME
DATE	DATA
VARCHAR	EMAIL

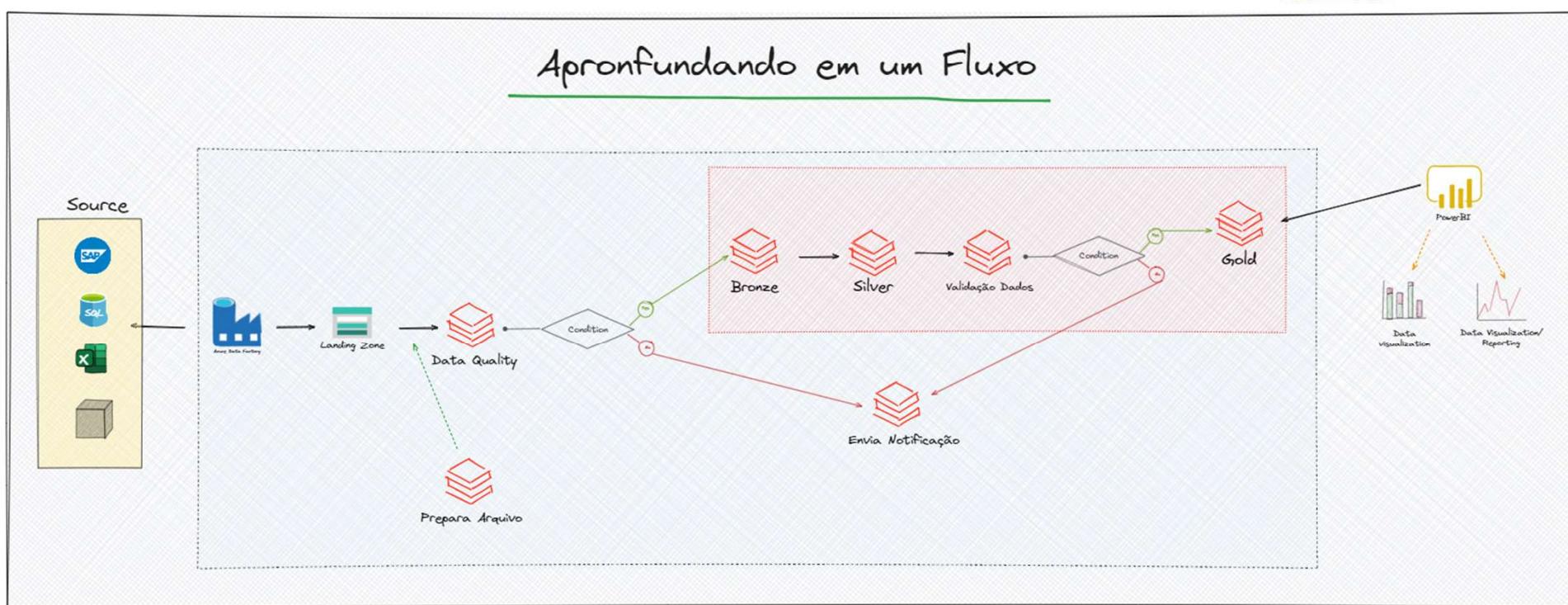
Como fazer isso funcionar?



Nova Arquitetura



Nova Arquitetura



Obrigado!



in

[gabriel-quintellao](#)



[dbaassists](#)



[quintellao](#)



<https://www.dbaassists.com.br/>

Referências/Dicas



- 1 – Faça muito Networking;
- 2 – Estude SQL independente da sua carreira! Quem trampa com dados TEM que saber muito SQL;
- 3 – Estude um Player de Nuvem (Azure, AWS ou GCP);
- 4 – Não acredite que irá conquistar a glória eterna + um super salário da noite pro dia, ou melhor, em 3 meses. É fria Bino!;
- 5 – Participe de Grupos do WhatsApp, Telegram, Discord;
- 6 – Produza materiais, é uma forma de estudar;
- 7 – Se possível procure um mentor para sua carreira, faça uma mentoria (claro, se possível! Nada de ficar endividado!)
- 8 – Valorize cursos da Udemy, Hotmart e outros...aqueles gráts também possuem valor, mesmo sendo gráts!
- 9 – Deixo algumas dicas de Profissionais:

Ítalo Mesquita - <https://www.linkedin.com/in/italomesquita/>

Projeto Evangelizando a Linguagem SQL - <https://luizlima.net/projeto-evangelizando-a-linguagem-sql/>

Raphael Amorim - <https://www.youtube.com/@bifastsolutions>

Wallace Camargo - <https://www.youtube.com/@wallacecamargo1043>

Luiz Lima (ou melhor a Família Lima) - <https://luizlima.net/> e <https://cursos.powertuning.com.br/?msg=not-logged-in>

Demetrius Mata - <https://www.linkedin.com/in/demetrius-mata-6aa74910a/?originalSubdomain=br>

Luciano Bolba - <https://www.youtube.com/@luhborba>

João Oliveira (Envolve Data) - <https://www.youtube.com/@joaomdeoliveira>

Consulta DB (Luiz Santana) - <https://www.youtube.com/@consultabd>

Téo Me Why - <https://www.twitch.tv/teomewhy/videos>

Dirceu Resende - <https://www.dirceuresende.com/blog/>

SQL WEEK RECIFE

Conference



O MAIOR EVENTO
ONLINE EM LINHA
RETA DO NORDESTE.