

GRUPO_E

SOLUTION SPRINT

PROPOSTA

Alexandre Henrique de Oliveira Parreira

Cristina Rodrigues Abrantes

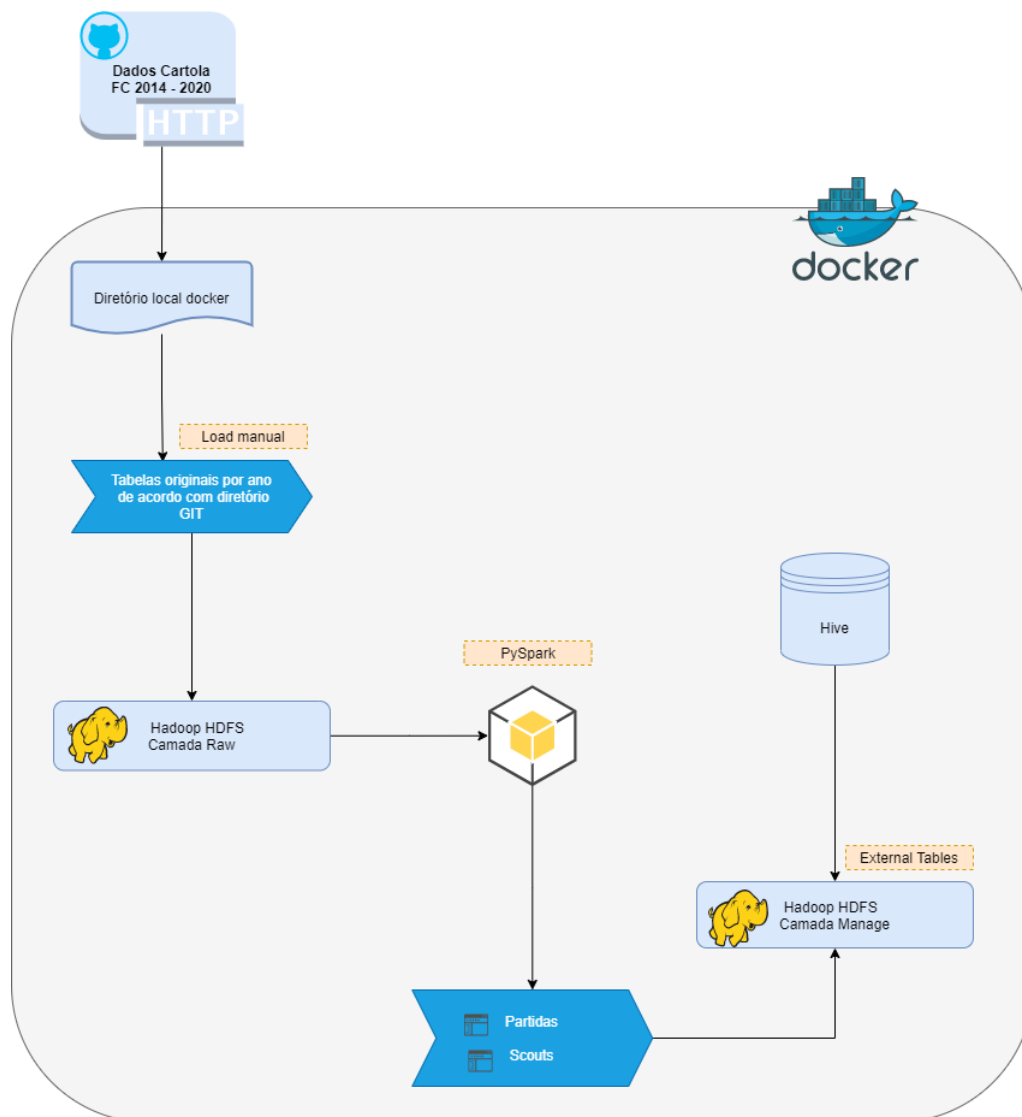
Gabriel Correia

Gustavo Santos Costa

TURMA 3ABDO

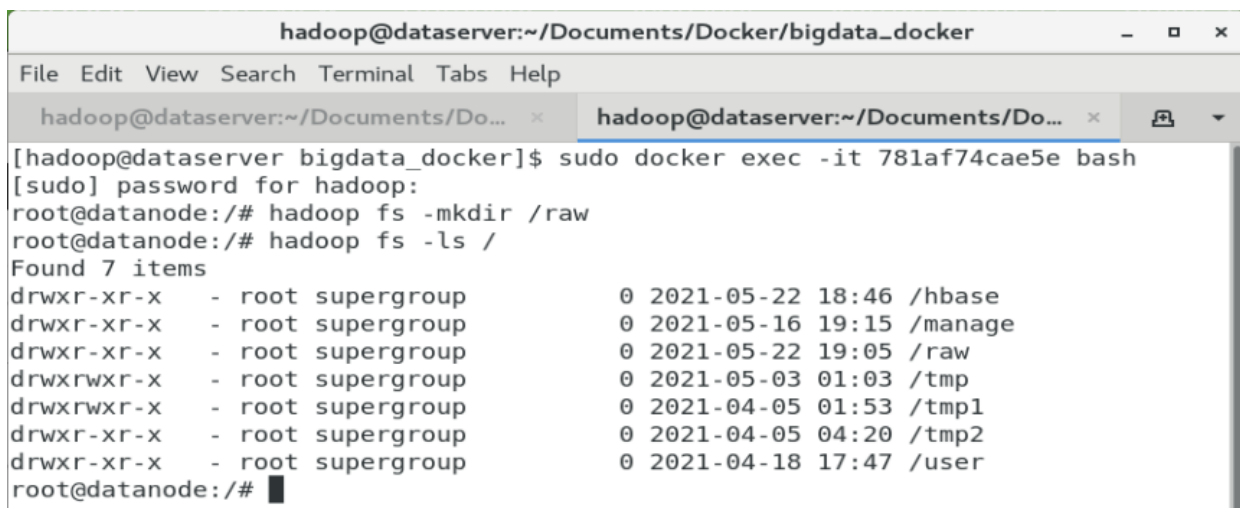
DESAFIO

1. Elaborar arquitetura para obtenção e tratamento de dados
2. Importar os arquivos para o HDFS
3. Criar tabelas no Hive para realizar consultas analíticas
4. Construir consultas SQL para responder às seguintes questões:
 - Quantos registros há na tabela por ano?
 - Quantas equipes únicas mandantes existem?
 - Quantas vezes as equipes mandantes saíram vitoriosas?
 - Quantas vezes as equipes visitantes saíram vitoriosas?
 - Quantas partidas resultaram em empate?
 - Quais jogadores detêm os melhores scouts gerais e por ano?
 - Qual é o time ideal?
 - Será que podemos preparar a ingestão para o campeonato de 2021?
 - Como capturar os dados direto do Cartola FC?



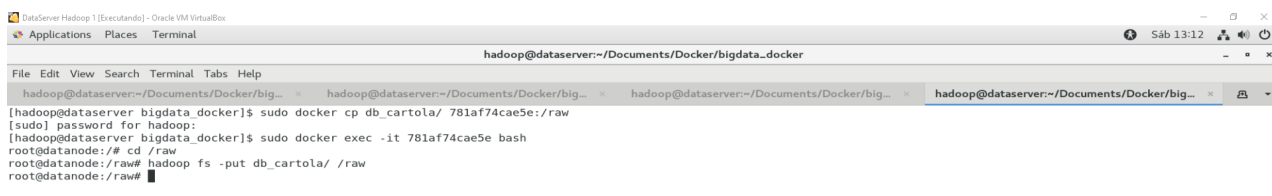
IMPORTAR OS ARQUIVOS PARA HDFS

- Criação de pasta no HDFS



```
hadoop@dataserver:~/Documents/Docker/bigdata_docker
File Edit View Search Terminal Tabs Help
hadoop@dataserver:~/Documents/Do... x hadoop@dataserver:~/Documents/Do... x
[hadoop@dataserver bigdata_docker]$ sudo docker exec -it 781af74cae5e bash
[sudo] password for hadoop:
root@datanode:/# hadoop fs -mkdir /raw
root@datanode:/# hadoop fs -ls /
Found 7 items
drwxr-xr-x - root supergroup 0 2021-05-22 18:46 /hbase
drwxr-xr-x - root supergroup 0 2021-05-16 19:15 /manage
drwxr-xr-x - root supergroup 0 2021-05-22 19:05 /raw
drwxrwxr-x - root supergroup 0 2021-05-03 01:03 /tmp
drwxrwxr-x - root supergroup 0 2021-04-05 01:53 /tmp1
drwxr-xr-x - root supergroup 0 2021-04-05 04:20 /tmp2
drwxr-xr-x - root supergroup 0 2021-04-18 17:47 /user
root@datanode:/#
```

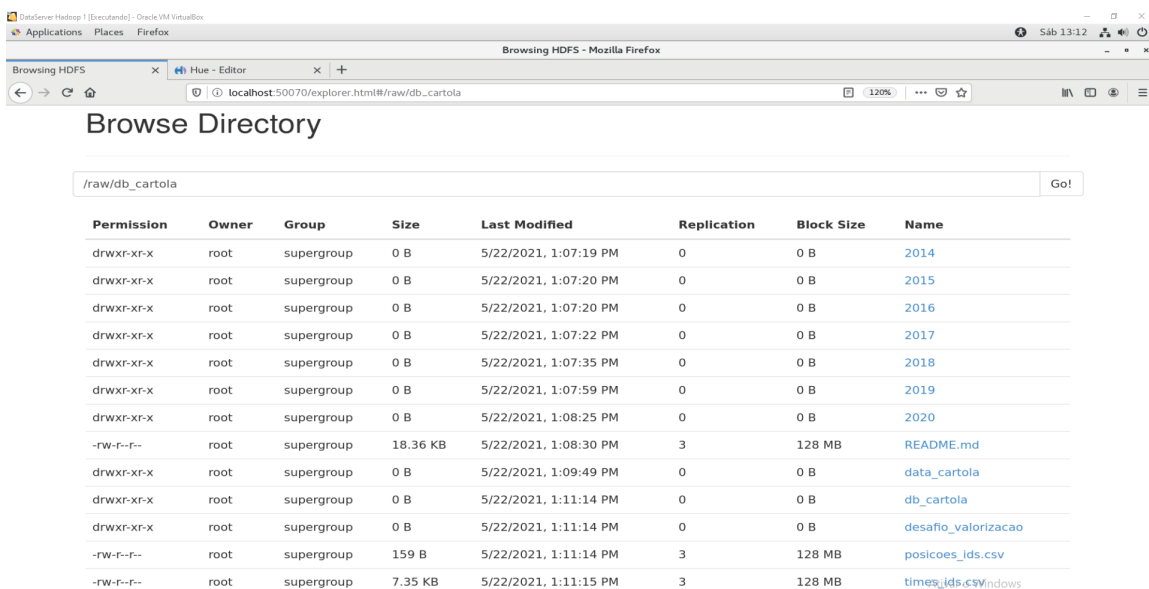
- Cópia dos arquivos .csv para o HDFS



```
dataServer Hadoop 1 [Executando] - Oracle VM VirtualBox
Applications Places Terminal
hadoop@dataserver:~/Documents/Docker/bigdata_docker
File Edit View Search Terminal Tabs Help
hadoop@dataserver:~/Documents/Docker/big... x hadoop@dataserver:~/Documents/Docker/big... x hadoop@dataserver:~/Documents/Docker/big... x hadoop@dataserver:~/Documents/Docker/big... x
[hadoop@dataserver bigdata_docker]$ sudo docker cp db_cartola/ 781af74cae5e:/raw
[sudo] password for hadoop:
[hadoop@dataserver bigdata_docker]$ sudo docker exec -it 781af74cae5e bash
root@datanode:/# cd /raw
root@datanode:/raw# hadoop fs -put db_cartola/ /raw
root@datanode:/raw#
```

IMPORTAR OS ARQUIVOS PARA HDFS

- Arquivos movidos para o HDFS



Browse Directory

/raw/db_cartola

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:19 PM	0	0 B	2014
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:20 PM	0	0 B	2015
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:20 PM	0	0 B	2016
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:22 PM	0	0 B	2017
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:35 PM	0	0 B	2018
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:07:59 PM	0	0 B	2019
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:08:25 PM	0	0 B	2020
-rw-r--r--	root	supergroup	18.36 KB	5/22/2021, 1:08:30 PM	3	128 MB	README.md
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:09:49 PM	0	0 B	data_cartola
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:11:14 PM	0	0 B	db_cartola
drwxr-xr-x	root	supergroup	0 B	5/22/2021, 1:11:14 PM	0	0 B	desafio_valorizacao
-rw-r--r--	root	supergroup	159 B	5/22/2021, 1:11:14 PM	3	128 MB	posicoes_ids.csv
-rw-r--r--	root	supergroup	7.35 KB	5/22/2021, 1:11:15 PM	3	128 MB	times_ids.csv

CRIAR TABELAS NO HIVE

```
CREATE EXTERNAL TABLE PARTIDAS_GERAL (  
  ID INT  
  ,RODADA INT  
  ,CASA STRING  
  ,VISITANTE STRING  
  ,PLACARCASA INT  
  ,PLACARVISITANTE INT  
  ,RESULTADO STRING  
  ,ANO INT  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
LOCATION '/MANAGE/PARTIDAS_GERAL'  
TBLPROPERTIES("SKIP.HEADER.LINE.COUNT"="1");
```

```
CREATE EXTERNAL TABLE SCOUTS_GERAL (  
  ATLETA STRING  
  ,RODADA INT  
  ,CLUBE STRING  
  ,POSICAO STRING  
  ,PONTOS FLOAT  
  ,PONTOSMEDIA FLOAT  
  ,PRECO FLOAT  
  ,PRECOVARIACAO FLOAT  
  ,ANO INT  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
LOCATION '/MANAGE/SCOUTS_GERAL'  
TBLPROPERTIES("SKIP.HEADER.LINE.COUNT"="1");
```

RESULTADO CONSULTAS SQL

01

Quantos registros há na tabela por ano?

Tabela de Partidas:





```
SELECT count(*) AS Qtd_Registros, Ano
FROM PARTIDAS_GERAL
GROUP BY Ano
ORDER BY Ano ASC
```

Query History		Saved Queries		Query Builder		Results (7)	
		qtd_registros				ano	
1	380					2014	
2	380					2015	
3	380					2016	
4	380					2017	
5	380					2018	
6	380					2019	
7	380					2020	

RESULTADO CONSULTAS SQL

Total de registros das tabelas Partidas e Scouts:





```
WITH CTE AS(  
SELECT COUNT(*) as QTD_REGISTROS,'Partidas' as  
Tabela FROM PARTIDAS_GERAL  
) , CTE2 AS (  
SELECT count(*) AS QTD_REGISTROS,'Scouts' as Tabela from scouts_geral  
)  
SELECT * FROM CTE  
UNION  
SELECT * FROM CTE2
```

Query History	Saved Queries	Query Builder	Results (2)
		_u1.qtd_registros	_u1.tabela
   	1	2660	Partidas
	2	173959	Scouts

02

Quantas equipes únicas mandantes existem?

```
SELECT count(distinct Casa) Qtde_Equipes_Unicas_Mandantes, ,Ano  
FROM PARTIDAS_GERAL  
GROUP BY Ano  
ORDER BY Ano ASC
```







Query History	Saved Queries	Query Builder	Results (7)
		equipes_mandantes	ano
   	1	20	2014
	2	20	2015
	3	20	2016
	4	20	2017
	5	20	2018
	6	20	2019
	7	20	2020

RESULTADO CONSULTAS SQL

03

Quantas vezes as equipes mandantes saíram vitoriosas?







```
SELECT count(*) as Qtde_Vitorias_Mandantes,Ano  
FROM PARTIDAS_GERAL  
WHERE Resultado = 'Casa'  
GROUP BY Ano  
ORDER BY Ano
```

Query History	Saved Queries	Query Builder	Results (7)
		qtde_vitorias_mandantes	ano
		1 197	2014
		2 194	2015
		3 202	2016
		4 167	2017
		5 201	2018
		6 184	2019
		7 171	2020

04

Quantas vezes as equipes visitantes saíram vitoriosas?

```
SELECT count(*) as Qtde_Vitorias_Visitante,Ano  
FROM PARTIDAS_GERAL  
WHERE Resultado = 'Visitante'  
GROUP BY Ano  
ORDER BY Ano
```

Query History	Saved Queries	Query Builder	Results (7)
		qtde_vitorias_visitante	ano
		1 91	2014
		2 80	2015
		3 83	2016
		4 110	2017
		5 67	2018
		6 98	2019
		7 101	2020

RESULTADO CONSULTAS SQL

05

Quantas partidas resultaram em empate?

```
SELECT count(*) as Qtde_Empates ,Ano
FROM PARTIDAS_GERAL
WHERE Resultado = 'Empate'
GROUP BY Ano
ORDER BY Ano
```

	Query History	Saved Queries	Query Builder	Results (7)
	qtde_empates			ano
1	92			2014
2	86			2015
3	94			2016
4	103			2017
5	112			2018
6	98			2019
7	108			2020

06

Quais jogadores detêm os melhores scouts gerais e por ano?

Scouts Gerais:

```
WITH cte as (
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,Ano desc) as colocacao ,round(sum(Pontos),2) as total_pontos,Atleta ,Clube,Ano
FROM scouts_geral
GROUP BY Atleta ,Clube,Ano)
SELECT * FROM cte WHERE colocacao = 1 ORDER BY total_pontos DESC
```

RESULTADO CONSULTAS SQL

Query History	Saved Queries	Query Builder	Results (7)		
	cte.colocacao	cte.total_pontos	cte.atleta	cte.clube	cte.ano
1	1	280.2	Vanderlei	SAN	2017
2	1	269.3	Gabriel	Flamengo	2019
3	1	254.7	Marinho	Santos	2020
4	1	233.6	Renê	FLA	2018
5	1	211	Marinho	Vitória	2016
6	1	203.2	Conca	fluminense	2014
7	1	201.1	Marcelo Lomba	Ponte Preta	2015

Melhores scouts de 2014:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,Ano desc)
as colocacao ,round(sum(Pontos),2) as total_pontos ,Atleta ,Clube,Ano
FROM scouts_geral
WHERE Ano = 2014
GROUP BY Atleta ,Clube,Ano
```

Query History	Saved Queries	Query Builder	Results (100+)		
	colocacao	total_pontos	atleta	clube	ano
1	1	203.2	Conca	fluminense	2014
2	2	201.3	Marcelo Grohe	grêmio	2014
3	3	186.6	Renan	goiás	2014
4	4	175.5	Marcelo Moreno	cruzeiro	2014
5	5	170	Ricardo Goulart	cruzeiro	2014
6	6	168.9	Rogério Ceni	são paulo	2014
7	7	168.8	Éverton Ribeiro	cruzeiro	2014
8	8	167.5	Fred	fluminense	2014
9	9	167.3	D'Alessandro	internacional	2014
10	10	163.4	Marcelo Oliveira	cruzeiro	2014

RESULTADO CONSULTAS SQL

Melhores scouts de 2015:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,Ano desc)
as colocacao ,round(sum(Pontos),2) as total_pontos,Atleta,Clube,Ano
FROM scouts_geral
WHERE Ano = 2015
GROUP BY Atleta ,Clube,Ano
```

Query History		Saved Queries		Query Builder		Results (100+)					
		colocacao		total_pontos		atleta		clube		ano	
		1	1	201.1		Marcelo Lomba		Ponte Preta		2015	
		2	2	189.9		Danilo Fernandes		Sport		2015	
		3	3	171.1		Jadson		Corinthians		2015	
		4	4	169.9		Cássio		Corinthians		2015	
		5	5	167.5		Luan		Grêmio		2015	
		6	6	165.98		Tite		Corinthians		2015	
		7	7	164.9		Rodrigo		Vasco		2015	
		8	8	162.1		Lucas Pratto		Atlético-MG		2015	
		9	9	157.8		Alexandre Pato		São Paulo		2015	
		10	10	157.4		Dudu		Palmeiras		2015	

Melhores scouts de 2016:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,
Ano desc) as colocacao
,round(sum(Pontos),2) as total_pontos,Atleta, ,Clube,Ano
FROM scouts_geral
WHERE Ano = 2016
GROUP BY Atleta ,Clube,Ano
```

Query History		Saved Queries		Query Builder		Results (100+)				
<div><div></div><div></div><div></div></div>		colocacao		total_pontos		atleta		clube		ano
<div><div></div><div></div><div></div></div>	1	1		211		Marinho		Vitória		2016
<div><div></div><div></div><div></div></div>	2	2		189.3		Vanderlei		Santos		2016
<div><div></div><div></div><div></div></div>	3	3		183.7		Arrascaeta		Cruzeiro		2016
	4	4		176.7		Jorge		Flamengo		2016
	5	5		171.8		Vitor Bueno		Santos		2016
	6	6		169.4		Keno		Santa Cruz		2016
	7	7		167		Diego Souza		Sport		2016
	8	8		164.3		Gustavo Scarpa		Fluminense		2016
	9	9		164.2		Wilson		Coritiba		2016
	10	10		157.9		Wellington Silva		Fluminense		2016

RESULTADO CONSULTAS SQL

Melhores scouts de 2017:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,  
Ano desc) as colocacao  
,round(sum(Pontos),2) as total_pontos,Atleta ,Clube,Ano  
FROM scouts_geral  
WHERE Ano = 2017  
GROUP BY Atleta ,Clube,Ano
```

	Query History	Saved Queries	Query Builder	Results (100+)		
					colocacao	total_pontos
						atleta
						clube
						ano
1	1				280.2	Vanderlei
2	2				221	Zé Rafael
3	3				203.3	Wilson
4	4				196.3	Lucca
5	5				182.5	Renê Júnior
6	6				179.2	Reinaldo
7	7				172.8	André
8	8				167.8	Douglas Friedrich
9	9				167.5	Arthur
10	10				164.9	Bruno Henrique

Melhores scouts de 2018:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos)  
desc,Ano desc) as colocacao ,round(sum(Pontos),2) as total_pontos ,Atleta  
,Clube,Ano  
FROM scouts_geral  
WHERE Ano = 2018  
GROUP BY Atleta ,Clube,Ano
```

	Query History	Saved Queries	Query Builder	Results (100+)		
					colocacao	total_pontos
						atleta
						clube
						ano
1	1				233.6	Renê
2	2				219.5	Yago Pikachu
3	3				212.7	Zé Rafael
4	4				208.7	Lucas Paquetá
5	5				199.7	Patrick
6	6				197.5	Victor Cuesta
7	7				197.4	Everton
8	8				196.7	Nico López
9	9				195.9	Dudu
10	10				194.8	Gabriel

RESULTADO CONSULTAS SQL

Melhores scouts de 2019:

```
SELECT  
  ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,Ano desc)  
as colocacao  
  ,round(sum(Pontos),2) as total_pontos ,Atleta ,Clube,Ano  
FROM scouts_geral  
WHERE Ano = 2019  
GROUP BY Atleta ,Clube,Ano
```

Query History		Saved Queries		Query Builder		Results (100+)					
		colocacao		total_pontos		atleta		clube		ano	
		1	1	269.3		Gabriel		Flamengo		2019	
		2	2	244.9		Bruno Henrique		Flamengo		2019	
		3	3	233.6		Tadeu		Goiás		2019	
		4	4	229.1		Arrascaeta		Flamengo		2019	
		5	5	214.1		Carlos Sánchez		Santos		2019	
		6	6	209.9		Everton		Grêmio		2019	
		7	7	200.4		Eduardo Sasha		Santos		2019	
		8	8	197		Jorge		Santos		2019	
		9	9	193.2		Gilberto		Bahia		2019	
		10	10	192.5		Reinaldo		São Paulo		2019	

Melhores scouts de 2020:

```
SELECT ROW_NUMBER() OVER (partition by Ano ORDER BY sum(Pontos) desc,Ano  
desc) as colocacao,round(sum(Pontos),2) as total_pontos,Atleta ,Clube,Ano  
FROM scouts_geral  
WHERE Ano = 2020  
GROUP BY Atleta ,Clube,Ano
```

Query History		Saved Queries		Query Builder		Results (100+)					
		colocacao		total_pontos		atleta		clube		ano	
		1	1	254.7		Marinho		Santos		2020	
		2	2	195.8		Thiago Galhardo		Internacional		2020	
		3	3	193.5		Vinicius		Ceará		2020	
		4	4	187.2		Keno		Atlético-MG		2020	
		5	5	172.8		Jean		Atlético-GO		2020	
		6	6	172.5		Guilherme Arana		Atlético-MG		2020	
		7	7	170.7		Luciano		São Paulo		2020	
		8	8	166		Pepê		Grêmio		2020	
		9	9	165.9		Patrick		Internacional		2020	
		10	10	154.3		Cano		Vasco		2020	

RESULTADO CONSULTAS SQL

07

Qual é o time ideal?

Time ideal de 2014:

```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao, Ano ORDER BY  
    avg(Pontos) DESC, Ano DESC) AS colocacao  
    , round(avg(Pontos), 2) AS total_pontos  
    , Atleta, Clube, Posicao, Ano  
  FROM scouts_geral  
  WHERE Ano = 2014  
  GROUP BY Atleta, Clube, Posicao, Ano)  
SELECT * FROM cte  
WHERE  
  (colocacao = 1 AND posicao LIKE 'Gol%')  
  OR (colocacao IN (1, 2) AND posicao LIKE 'Lat%') OR (colocacao IN (1, 2) AND  
  posicao LIKE 'Zag%') OR (colocacao IN (1, 2, 3) AND posicao LIKE 'Mei%') OR  
  (colocacao IN (1, 2, 3) AND posicao LIKE 'Ata%') OR (colocacao IN (1) AND posicao  
  LIKE 'Téc%')  
ORDER BY colocacao ASC
```

	cte.colocacao	cte.total_pontos	cte.atleta	cte.clube	cte.posicao	cte.ano
1	1	4.71	Cléber	corinthians	Zagueiro	2014
2	1	4.3	Marcelo Oliveira	cruzeiro	Técnico	2014
3	1	6.3	Ricardo Goulart	cruzeiro	Meia	2014
4	1	9.8	Ronan	fluminense	Lateral	2014
5	1	5.59	Marcelo Grohe	grêmio	Goleiro	2014
6	1	6.9	Alan Kardec	palmeiras	Atacante	2014
7	2	4.62	Dedé	cruzeiro	Zagueiro	2014
8	2	5.49	Conca	fluminense	Meia	2014
9	2	6.1	Wellington Silva	internacional	Lateral	2014
10	2	6.2	Marquinhos	vitória	Atacante	2014
11	3	5.45	Éverton Ribeiro	cruzeiro	Meia	2014
12	3	5.78	Fred	fluminense	Atacante	2014

RESULTADO CONSULTAS SQL

Time ideal de 2015:





```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao, Ano ORDER BY  
    avg(Pontos) DESC, Ano DESC) AS colocacao  
    , round(avg(Pontos), 2) AS total_pontos  
    , Atleta, Clube, Posicao, Ano  
  FROM scouts_geral  
  WHERE Ano = 2015  
  GROUP BY Atleta, Clube, Posicao, Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'Gol%')  
OR (colocacao IN (1, 2) AND posicao LIKE 'Lat%') OR (colocacao IN (1, 2) AND  
posicao LIKE 'Zag%')  
OR (colocacao IN (1, 2, 3) AND posicao LIKE 'Mei%') OR (colocacao IN (1, 2, 3) AND  
posicao LIKE 'Ata%')  
OR (colocacao IN (1) AND posicao LIKE 'Téc%') ORDER BY colocacao ASC
```

Query History		Saved Queries		Query Builder		Results (12)					
						cte.colocacao	cte.total_pontos	cte.atleta	cte.clube	cte.posicao	cte.ano
<div><div></div><div></div><div></div></div>	1	1				1	4.58	Rodrigo	Vasco	Zagueiro	2015
	2	1				2	8.45	Wanderley Filho	Goiás	Técnico	2015
	3	1				3	5	Pedro Ken	Corinthians	Meia	2015
	4	1				4	5.4	Matias Rodriguez	Grêmio	Lateral	2015
	5	1				5	5.59	Marcelo Lomba	Ponte Preta	Goleiro	2015
	6	1				6	8.9	Rafhael Lucas	Corinthians	Atacante	2015
	7	2				7	3.71	Paulo Miranda	São Paulo	Zagueiro	2015
	8	2				8	4.83	Renato Cajá	Ponte Preta	Meia	2015
	9	2				9	4.21	Fábio Santos	Corinthians	Lateral	2015
	10	2				10	5.8	Nenê	Vasco	Atacante	2015
	11	3				11	4.75	Jadson	Corinthians	Meia	2015
	12	3				12	5.54	Joelinton	Sport	Atacante	2015

RESULTADO CONSULTAS SQL

Time ideal de 2016:




```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao,Ano ORDER BY avg(Pontos)  
    DESC,Ano DESC) AS colocacao  
    ,round(avg(Pontos),2) AS total_pontos  
    ,Atleta  
    ,Clube  
    ,Posicao  
    ,Ano  
  FROM scouts_geral  
  WHERE Ano = 2016  
  GROUP BY Atleta ,Clube,Posicao,Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'Gol%')  
OR (colocacao IN (1,2) AND posicao LIKE 'Lat%')  
OR (colocacao IN (1,2) AND posicao LIKE 'Zag%')  
OR (colocacao IN(1,2,3) AND posicao LIKE 'Mei%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'Ata%')  
OR (colocacao IN (1) AND posicao LIKE 'Téc%')  
ORDER BY colocacao ASC
```

Query History		Saved Queries		Query Builder		Results (12)								
		cte.colocacao		cte.total_pontos		cte.atleta		cte.clube		cte.posicao		cte.ano		
		1	1			5.78		Felipe		Corinthians		Zagueiro		2016
		2	1			5.47		Muricy Ramalho		Flamengo		Técnico		2016
		3	1			5.91		Bruno Henrique		Corinthians		Meia		2016
		4	1			4.65		Jorge		Flamengo		Lateral		2016
		5	1			4.98		Vanderlei		Santos		Goleiro		2016
		6	1			5.55		Marinho		Vitória		Atacante		2016
		7	2			4.05		Pedro Geromel		Grêmio		Zagueiro		2016
		8	2			5.24		Giuliano		Grêmio		Meia		2016
		9	2			4.19		Alemão		Botafogo		Lateral		2016
		10	2			4.63		Copete		Santos		Atacante		2016
		11	3			4.83		Arrascaeta		Cruzeiro		Meia		2016
		12	3			4.49		Andres Chavez		São Paulo		Atacante		2016

RESULTADO CONSULTAS SQL

Time ideal de 2017:




```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao,Ano ORDER BY  
    avg(Pontos) DESC,  
    Ano DESC) AS colocacao ,round(avg(Pontos),2) AS total_pontos ,Atleta,Clube  
    ,Posicao ,Ano  
  FROM scouts_geral  
  WHERE Ano = 2017  
  GROUP BY Atleta ,Clube,Posicao,Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'gol%')  
OR (colocacao IN (1,2) AND posicao LIKE 'lat%')  
OR (colocacao IN (1,2) AND posicao LIKE 'zag%')  
OR (colocacao IN(1,2,3) AND posicao LIKE 'mei%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'ata%')  
OR (colocacao IN (1) AND posicao LIKE 'tec%')  
ORDER BY colocacao ASC
```

Query History		Saved Queries		Query Builder		Results (12)				
		cte.colocacao		cte.total_pontos		cte.atleta		cte.clube	cte.posicao	cte.ano
		1	1	4.23		Baibuena		COR	zag	2017
		2	1	4.6		Alberto Valentim		PAL	tec	2017
		3	1	6.13		Hernanes		SAO	mei	2017
		4	1	4.72		Reinaldo		CHA	lat	2017
		5	1	7.37		Vanderlei		SAN	gol	2017
		6	1	5.17		Lucca		PON	ata	2017
		7	2	3.55		Wallace Reis		VIT	zag	2017
		8	2	5.82		Zé Rafael		BAH	mei	2017
		9	2	4.17		Sidcley		ATL	lat	2017
		10	2	5.05		Luiz Araújo		SAO	ata	2017
		11	3	4.8		René Júnior		BAH	mei	2017
		12	3	4.69		Richarlison		FLU	ata	2017

RESULTADO CONSULTAS SQL

Time ideal de 2018:




```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao,Ano ORDER BY  
    avg(Pontos) DESC,Ano DESC) AS colocacao  
    ,round(avg(Pontos),2) AS total_pontos  
    ,Atleta  
    ,Clube  
    ,Posicao  
    ,Ano  
  FROM scouts_geral  
  WHERE Ano = 2018  
  GROUP BY Atleta ,Clube,Posicao,Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'gol%')  
OR (colocacao IN (1,2) AND posicao LIKE 'lat%')  
OR (colocacao IN (1,2) AND posicao LIKE 'zag%')  
OR (colocacao IN(1,2,3) AND posicao LIKE 'mei%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'ata%')  
OR (colocacao IN (1) AND posicao LIKE 'tec%')  
ORDER BY colocacao ASC
```

Query History	Saved Queries	Query Builder	Results (12)			
	cte.colocacao	cte.total_pontos	cte.atleta	cte.clube	cte.posicao	cte.ano
	1	5.2	Victor Cuesta	INT	zag	2018
	2	5.27	Luiz Felipe Scolari	PAL	tec	2018
	3	8.13	Anselmo	SPO	mei	2018
	4	7.32	Thiago Carleto	ATL	lat	2018
	5	4.96	Éverson	CEA	gol	2018
	6	10.63	Róger Guedes	ATL	ata	2018
	7	4.45	Balbuena	COR	zag	2018
	8	7.12	Otero	ATL	mei	2018
	9	6.15	René	FLA	lat	2018
	10	6.61	Vinicius Junior	FLA	ata	2018
	11	5.6	Zé Rafael	BAH	mei	2018
	12	5.81	Vitinho	FLA	ata	2018

RESULTADO CONSULTAS SQL

Time ideal de 2019:

```
WITH cte AS (  
  SELECT ROW_NUMBER() OVER (PARTITION BY Posicao,Ano ORDER BY avg(Pontos)  
    DESC,Ano DESC) AS colocacao  
    ,round(avg(Pontos),2) AS total_pontos  
    ,Atleta  
    ,Clube  
    ,Posicao  
    ,Ano  
  FROM scouts_geral  
  WHERE Ano = 2019  
  GROUP BY Atleta ,Clube,Posicao,Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'gol%')  
OR (colocacao IN (1,2) AND posicao LIKE 'lat%')  
OR (colocacao IN (1,2) AND posicao LIKE 'zag%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'mei%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'ata%')  
OR (colocacao IN (1) AND posicao LIKE 'tec%')  
ORDER BY colocacao ASC
```

Query History		Saved Queries		Query Builder		Results (11)				
				cte.colocacao	cte.total_pontos	cte.atleta	cte.clube	cte.posicao	cte.ano	
				1	1	4.86	Victor Cuesta	Internacional	zag	2019
				2	1	6.03	Arrascaeta	Flamengo	mei	2019
				3	1	5.73	Ramon Pereira	Avaí	lat	2019
				4	1	6.15	Tadeu	Goiás	gol	2019
				5	1	7.09	Gabriel	Flamengo	ata	2019
				6	2	3.93	Rafael Vaz	Goiás	zag	2019
				7	2	5.63	Carlos Sánchez	Santos	mei	2019
				8	2	5.18	Jorge	Santos	lat	2019
				9	2	6.44	Bruno Henrique	Flamengo	ata	2019
				10	3	4.39	Thiago Galhardo	Ceará	mei	2019
				11	3	5.52	Everton	Grêmio	ata	2019

RESULTADO CONSULTAS SQL

Time ideal de 2020:

```
WITH cte AS (  
SELECT ROW_NUMBER() OVER (PARTITION BY Posicao,Ano ORDER BY avg(Pontos)  
DESC,  
Ano DESC) AS colocacao,round(avg(Pontos),2) AS total_pontos ,Atleta ,Clube  
,Posicao ,Ano  
FROM scouts_geral  
WHERE Ano = 2020  
GROUP BY Atleta ,Clube,Posicao,Ano)  
SELECT * FROM cte  
WHERE  
(colocacao = 1 AND posicao LIKE 'gol%')  
OR (colocacao IN (1,2) AND posicao LIKE 'lat%')  
OR (colocacao IN (1,2) AND posicao LIKE 'zag%')  
OR (colocacao IN(1,2,3) AND posicao LIKE 'mei%')  
OR (colocacao IN (1,2,3) AND posicao LIKE 'ata%')  
OR (colocacao IN (1) AND posicao LIKE 'tec%')  
ORDER BY colocacao ASC
```

Query History		Saved Queries		Query Builder		Results (11)				
		cte.colocacao		cte.total_pontos		cte.atleta		cte.clube	cte.posicao	cte.ano
1	1			3.27		Bruno Fuchs		Internacional	zag	2020
2	1			5.15		Thiago Galhardo		Internacional	mei	2020
3	1			4.54		Guilherme Arana		Atlético-MG	lat	2020
4	1			4.55		Jean		Atlético-GO	gol	2020
5	1			6.88		Marinho		Santos	ata	2020
6	2			2.91		Sabino		Coritiba	zag	2020
7	2			5.09		Vinicius		Ceará	mei	2020
8	2			3.49		Guga		Atlético-MG	lat	2020
9	2			4.93		Keno		Atlético-MG	ata	2020
10	3			4.48		Patrick		Internacional	mei	2020
11	3			4.88		Luciano		São Paulo	ata	2020

RESULTADO CONSULTAS SQL

08

Será que podemos preparar a ingestão para o campeonato de 2021?

Sim, a nossa solução está preparada para receber os dados de 2021 pois a nossa forma de tratar, armazenar e disponibilizar esses dados em arquivos gerais na camada Manage suporta facilmente a entrada de um novo ano e facilita análises comparativas dos anos.

09

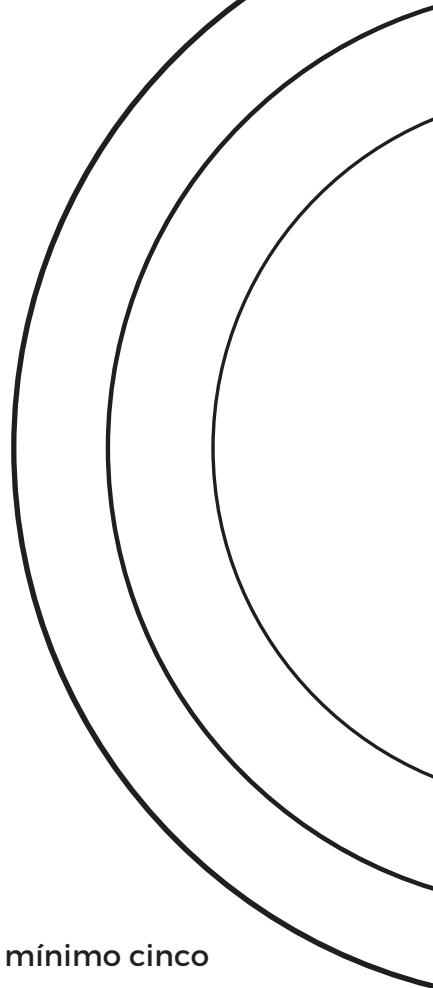
Será que podemos preparar a ingestão para o campeonato de 2021?

Podemos capturar acessando a API do Cartola FC, mas devido há um problema nesse serviço atualmente estamos capturando os dados do repositório do Git.

DESAFIO

Elaborar uma visão sobre as principais ferramentas (mínimo três) para governança de dados existentes no mercado. Esse documento deve conter:

- Breve descrição de cada ferramenta pesquisada e analisada.
- Quadro comparativo entre as ferramentas pesquisadas com no mínimo cinco características destacadas.
- Recomendação sobre o uso da ferramenta associado às boas práticas de governança.



BREVE DESCRIÇÃO DAS FERRAMENTAS PESQUISADAS

COLLIBRA

Aceleramos resultados de negócios confiáveis conectando os dados, percepções e algoritmos certos para todos os Data Citizens

INFORMATICA

Oferece uma verdadeira solução de governança de dados corporativos que pode ser usada localmente ou na nuvem, com dados tradicionais ou Big Data, para atender às necessidades da empresa e do setor de TI.

DENODO

Denodo é líder em virtualização de dados, fornecendo acesso a dados, governança de dados e recursos de entrega de dados em uma ampla gama de fontes de dados corporativos, em nuvem, big data e não estruturados, sem mover os dados de seus repositórios originais.

ALATION

Uma plataforma baseada em IA para pesquisa e descoberta de dados, governança de dados, administração de dados, análise e transformação digital.

APACHE ATLAS

O Atlas é um conjunto escalonável e extensível de serviços básicos de governança - permitindo que as empresas atendam com eficácia e eficiência seus requisitos de conformidade dentro do Hadoop e permite a integração com todo o ecossistema de dados corporativos.

O Apache Atlas fornece recursos abertos de gerenciamento e governança de metadados para que as organizações criem um catálogo de seus ativos de dados, classifiquem e controlem esses ativos e forneçam recursos de colaboração em torno desses ativos de dados para cientistas de dados, analistas e a equipe de governança de dados.

QUADRO COMPARATIVO

Funcionalidade	Collibra	Informatica	Denodo	Alation	Atlas
Armazenamento de Operações de Dados	-	-	-	-	-
Arquitetura de Dados	-	-	-	-	-
Dados mestres e de referências	-	X	-	-	-
Datawarehousing e Inteligência de Negócios	-	-	-	-	-
Gerenciamento de Documentos e Conteúdo	X	-	-	X	-
Gestão de Metadados	X	X	X	X	X
Glossário de Negócio	X	-	-	X	-
Integração de Dados e Interoperabilidade	-	X	X	-	-
Linhagem de Dados	X	-	X	-	X
Modelagem e Design de Dados	-	-	-	-	-
Política de Dados	X	X	X	-	-
Qualidade de Dados	X	X	-	X	-
Segurança de Dados	X	X	X	-	X

Através do <https://www.g2.com/compare/informatica-cloud-data-quality-vs-collibra-vs-denodo-vs-alation> podemos ver também uma comparação das ferramentas Collibra, Informatica, Denodo e Alation pelo site G2.

FUNCIONALIDADES ANALISADAS

- **Armazenamento de Operações de Dados:** Implementação e gestão do armazenamento de ativos de dados físicos estruturados (conhecido como Operação de Dados na primeira edição do DAMA-DMBOK)
- **Arquitetura de Dados:** A estrutura geral de dados e recursos relacionados a dados como parte integrante da arquitetura corporativa
- **Dados mestres e de referências:** Gerenciando dados compartilhados para reduzir a redundância e garantir uma melhor qualidade de dados por meio da definição padronizada e do uso de valores de dados
- **Datawarehousing e Inteligência de Negócios:** Gestão do processamento de dados analíticos e permissão do acesso a dados de suporte para relatórios e análises
- **Gerenciamento de Documentos e Conteúdo:** Armazenamento, proteção, indexação e acesso a dados encontrados em fontes não estruturadas (arquivos eletrônicos e registros físicos), disponibilizando esses dados para integração e interoperabilidade com dados estruturados (bancos de dados)
- **Gestão de Metadados:** Coleta, categorização, manutenção, integração, controle, gerenciamento e distribuição de metadados
- **Integração de Dados e Interoperabilidade:** Aquisição, extração, transformação, movimentação, entrega, replicação, federação, virtualização e suporte operacional (nova área de conhecimento no DMBOK2)
- **Modelagem e Design de Dados:** A modelagem de dados é o processo de descoberta, análise e determinação do escopo dos requisitos de dados e, em seguida, sua representação e comunicação de uma maneira precisa chamada de modelo de dados.

FUNCIONALIDADES ANALISADAS

- **Segurança de Dados:** Garantir privacidade, confidencialidade e acesso adequado [aos dados]
- **Política de Dados:** Operacionalize e gerencie políticas em todo o ciclo de vida da privacidade e dimensione a conformidade em novas regulamentações.
- **Glossário de Negócio:** Descubra e entenda os dados importantes para que você possa gerar insights que gerem valor comercial.
- **Linhagem de Dados:** Mostra como os dados fluem de sistema para sistema, com visualização de linhagem completa de ponta a ponta.
- **Qualidade de Dados:** Definição, monitoramento e manutenção da integridade dos dados e melhoria da qualidade dos dados

RECOMENDAÇÃO

Collibra é recomendada para as atividades de manutenção da qualidade dos dados, uma vez que sua plataforma é capaz de gerar automaticamente políticas de qualidade sobre os dados. Ela faz isso agrupando várias fontes de dados, em seguida executa um modelo de Machine Learning (ML) escolhido pelo usuário. Assim, o algoritmo OwIDQ consegue detectar dados duplicados, padrões categóricos de coluna cruzada e análise de outliers.

Além disso, a plataforma da Collibra é amigável tanto para equipe de TI quanto para equipes de negócio. Logo, tanto pessoas com formação técnica avançada na área quanto pessoas sem formação são capazes de utilizar a ferramenta. Ambas podem contribuir para cultura data-driven e inserção da cultura de governança.

Com relação à segurança dos dados, é possível atribuir cargos e respectivas permissões de uso dos dados. Dessa forma, permite-se o uso apenas dos dados relevantes para cada área.

Informatica possui várias soluções independentes, logo os clientes podem contratar conforme sua necessidade. É capaz de auxiliar na geração do catálogo de dados por meio de Inteligência Artificial (IA) e ML. Com esse catálogo, é possível unificar os metadados e suas respectivas definições (contexto). A Informatica pode ajudar empresas que ainda não possuem ou precisam de engenharia de dados com seus produtos focados nessa especialidade. Além disso, a ferramenta permite a ingestão e análise de big data de fontes como multi-cloud, hybrid ou on-premises, inclusive na modalidade de streaming.

Os produtos de Data Quality e Governança dessa ferramenta permitem a colaboração das equipes TI e negócio, assegura a qualidade dos dados independentemente do volume ou tipo e ainda conta com enriquecimento de base de clientes (verificar se é legal perante LGPD).

Com relação à segurança dos dados, a Informatica é capaz de identificar continuamente riscos e sugerir ações, registrar os acessos aos dados pelos usuários e controle de permissões de acesso. E ainda, ela auxilia no entendimento completo dos dados confidenciais e a exposição deles ao risco, precavendo a CCPA e GDPR.

Mais uma recomendação, é que a Informatica possui uma solução escalável de MDM.

Denodo possui um catálogo dinâmico de dados. Aceita diferentes tipos de fontes. Um item que é destacado pelos usuários é a sua capacidade de Data Virtualization, em que a manipulação do dado é facilitada e sua consulta também.

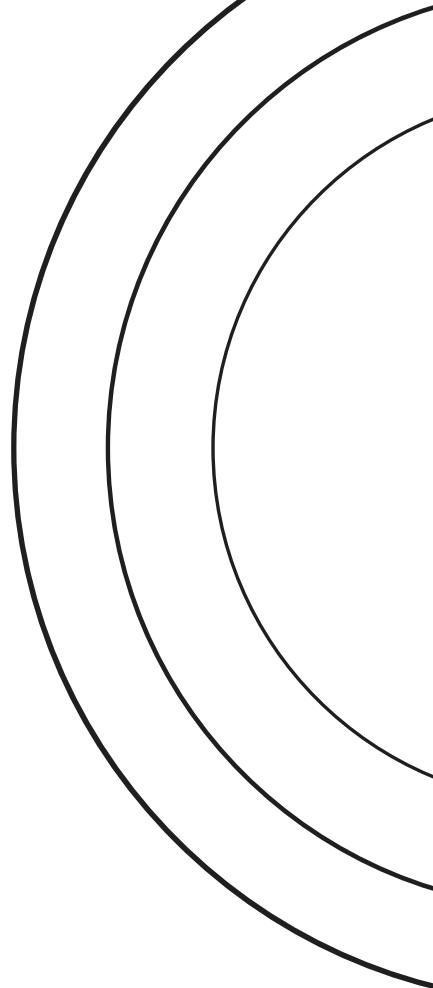
A Alation é destacada pelo desempenho ao gerenciar documentos e conteúdos de seus clientes.

O Apache Atlas é amplamente conhecido como uma ferramenta de Data Governance e ainda é uma ferramenta gratuita. Com suas limitações, é recomendada para empresas com projeto de governança ainda não tão complexo.

DESAFIO

Uma organização decidiu seguir as boas práticas sugeridas pela Governança de Dados e investir em Master Data Management. Assim precisamos detalhar as seguintes atividades:

- Elaborar Scrum Product Backlog.
- Propor planejamento do Product Backlog.
- Propor planejamento do Sprint.



BACKLOG INICIAL

- avaliar a necessidade do negócio / levantamento de requisitos
- mapear grau de maturidade atual quanto a governança e uso de dados mestres
- identificar fatores críticos para sucesso e stakeholders do projeto de implantação e do programa
- definir os objetivos iniciais da implantação do MDM (priorização cliente, fornecedores, parceiros)
- formulação de estrutura de governança responsável pelo MDM dentro da empresa (equipes e workflow)
- aprovação por parte dos patrocinadores dos casos definidos como prioritários
- identificar formas de incorporar a Estratégia MDM na implantação e manutenção de projetos
- desenvolver framework para padronização de datasets e identificar conflitos que precisarão resolução [critérios para avaliação do dado]
- identificar os data owners dos principais repositórios de dados
- em parceria com os data owners identificar datasets com alto valor de acordo com a priorização e repositório atual
- definir regras e políticas para criação de novos dados mestres (guidelines)
- data discovery e data matching em alto nível
- avaliação de tecnologia (POCS)
- definições de tecnologias
- desenvolver plano de melhoria contínua de Data Quality
- propor arquitetura lógica de dados e técnica detalhada To Be
- definir regras para padronização
- definir regras para particionamento
- definir regras para lineage
- definir regras para glossário de negócio
- criar scripts para ingestão dos dados (pipeline)
- migrar soluções de analytics para consumo a partir do SOT
- medir resultados das novas análises a partir dos dados mestres
- medir engajamento
- medir qualidade de dados mestres
- verificar causa de problemas de engajamento e no programa de qualidade de dados
- atualizar plano de melhoria contínua de Data Quality
- executar plano de melhoria
- testar e validar soluções migradas junto aos data owners e negócio

ROADMAP - IMPLANTAÇÃO MDM

1ª Release - discovery

- 1ª sprint da 1ª release
- Lead time: 20 dias corridos

- avaliar a necessidade do negócio / levantamento de requisitos
- mapear grau de maturidade atual quanto a governança e uso de dados mestres
- identificar fatores críticos para sucesso e stakeholders do projeto de implantação e do programa

- definir os objetivos iniciais da implantação do MDM (priorização cliente, fornecedores, parceiros)
- aprovação por parte dos patrocinador dos cases definidos como prioritários
- formulação de estrutura de governança responsável pelo MDM dentro da empresa (equipes e workflow)

1ª Release - acordos

- 2ª sprint da 1ª release
- Lead time: 40 dias corridos

1ª Release - estratégia

- 3ª sprint da 1ª release
- Lead time: 60 dias corridos

- identificar o data owners dos principais repositórios de dados
- em parceria com os data owners identificar datasets com alto valor de acordo com a priorização e repositório atual
- identificar e documentar formas de incorporar a Estratégia MDM na implantação e manutenção de projetos

ROADMAP - IMPLANTAÇÃO MDM

1ª Release - hands-on

- 4ª sprint da 1ª release
- Lead time: 80 dias corridos

- desenvolver framework para padronização de datasets e identificar conflitos que precisarão resolução [critérios para avaliação do dado]
- data discovery e data matching em alto nível (manual)
- propor de arquitetura lógica de dados e técnica detalhada To Be
- avaliação de tecnologia (pocs)

- definir regras e politicas para criação de dados mestres (guidelines)
- desenvolver plano de melhoria contínua de Data Quality
- definição de arquitetura alvo
- definição de soluções tecnológicas

2ª Release - acordos

- 1ª sprint da 2ª release
- Lead time: 100 dias corridos

2ª Release - SOT

- 2ª sprint da 2ª release
- Lead time: 120 dias corridos

- definir regras para padronização
- definir regras para particionamento
- definir regras para lineage
- definir regras para glossário de negócio

ROADMAP - IMPLANTAÇÃO MDM

2ª Release - Produção

- 3ª sprint da 2ª release
- Lead time: 140 dias corridos

- criar scripts para ingestão dos dados (pipeline)
- migrar soluções de analytics para consumo a partir do SOT
- testar e validar soluções migradas junto aos data owners e negócio

- medir resultados das novas análises a partir dos dados mestres
- medir em engajamento
- medir qualidade de dados mestres
- atualizar plano de melhoria contínua de Data Quality
- executar plano de melhoria

2ª Release - checkpoint

- 4ª sprint da 2ª release
- Lead time: 160 dias corridos

Proposta de backlog

Muitas das histórias listadas na 1ª versão do backlog podem ser classificadas como épicas.

Todas devem ser refinadas durante o planejamento de das sprints.

REFERÊNCIA

Case study – Master Data Management - Fujitsu

Best practices for a Successful MDM Implementario - Infosys

CHHS Master Data Management Strategy

<https://chhsdata.github.io/dataplaybook/documents/CHHS-Master-Data-Management-Strategy.pdf>

Trusted data for your entire organization | Collibra

Governança de dados corporativos: conformidade holística em escala | Informatica Brasil

Denodo, the Leader in Data Virtualization

Enterprise Data Catalog & Data Governance | Alation