Estudos de Caso em Big Data na SEFAZ-MT

Guilherme Falcão da Silva Campos

SEFAZ
SECRETARIA DE ESTADO DE FAZENDA

GOVERNO DE MATO GROSSO
ESTADO DE TRANSFORMAÇÃO

Por que Big Data?

- Criação de relatórios transacionais
- Hoje existem 1.143 relatórios
- Dificuldade na criação de modelos tradicionais de BI
- Alguns relatórios não executam em tempo hábil

Ambiente Transacional

ORACLE
EXADATA

X2-2 1/2

Storage
X5 1/4

X

DELL
cloudera

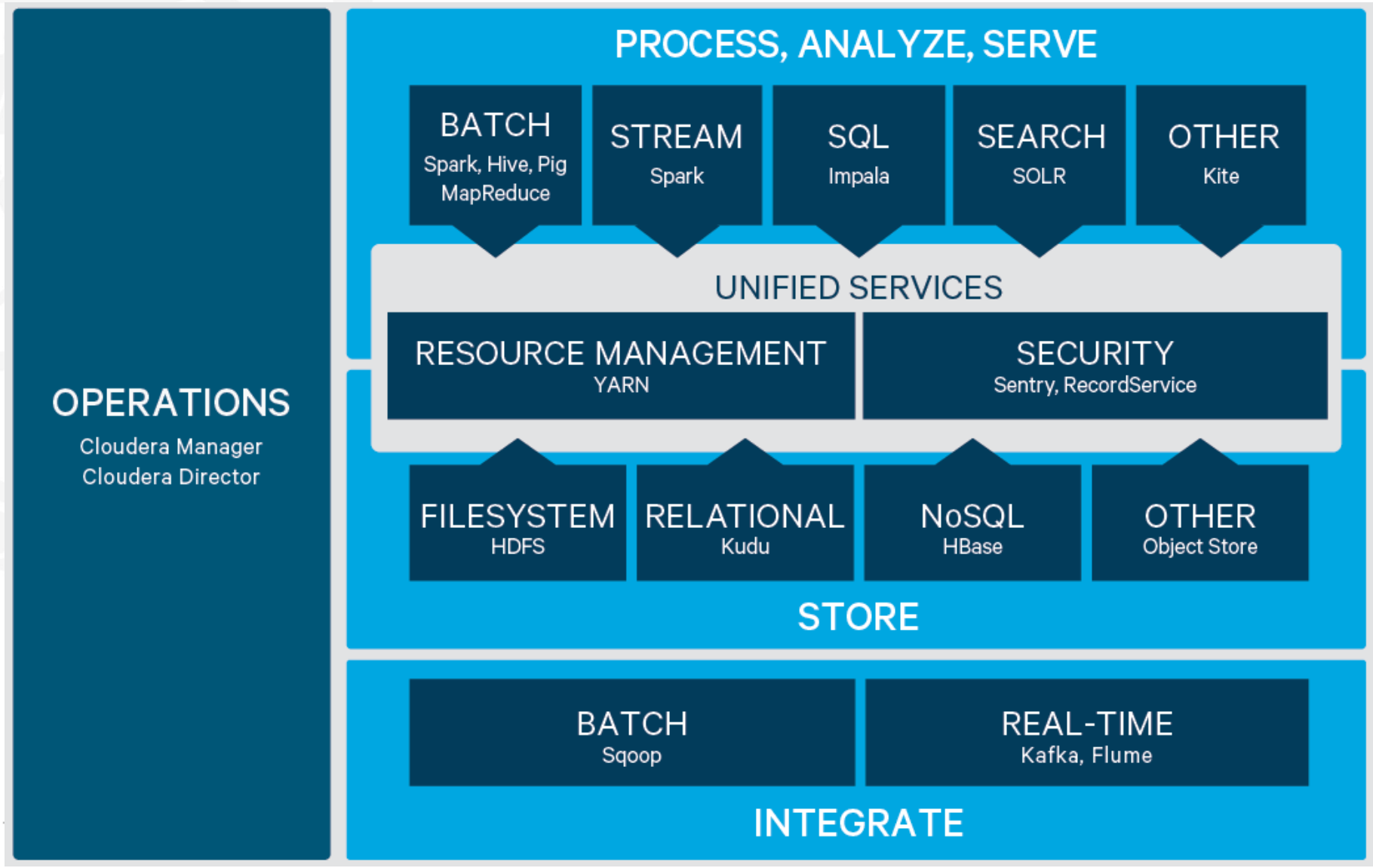
10x R730

Ambiente Cloudera

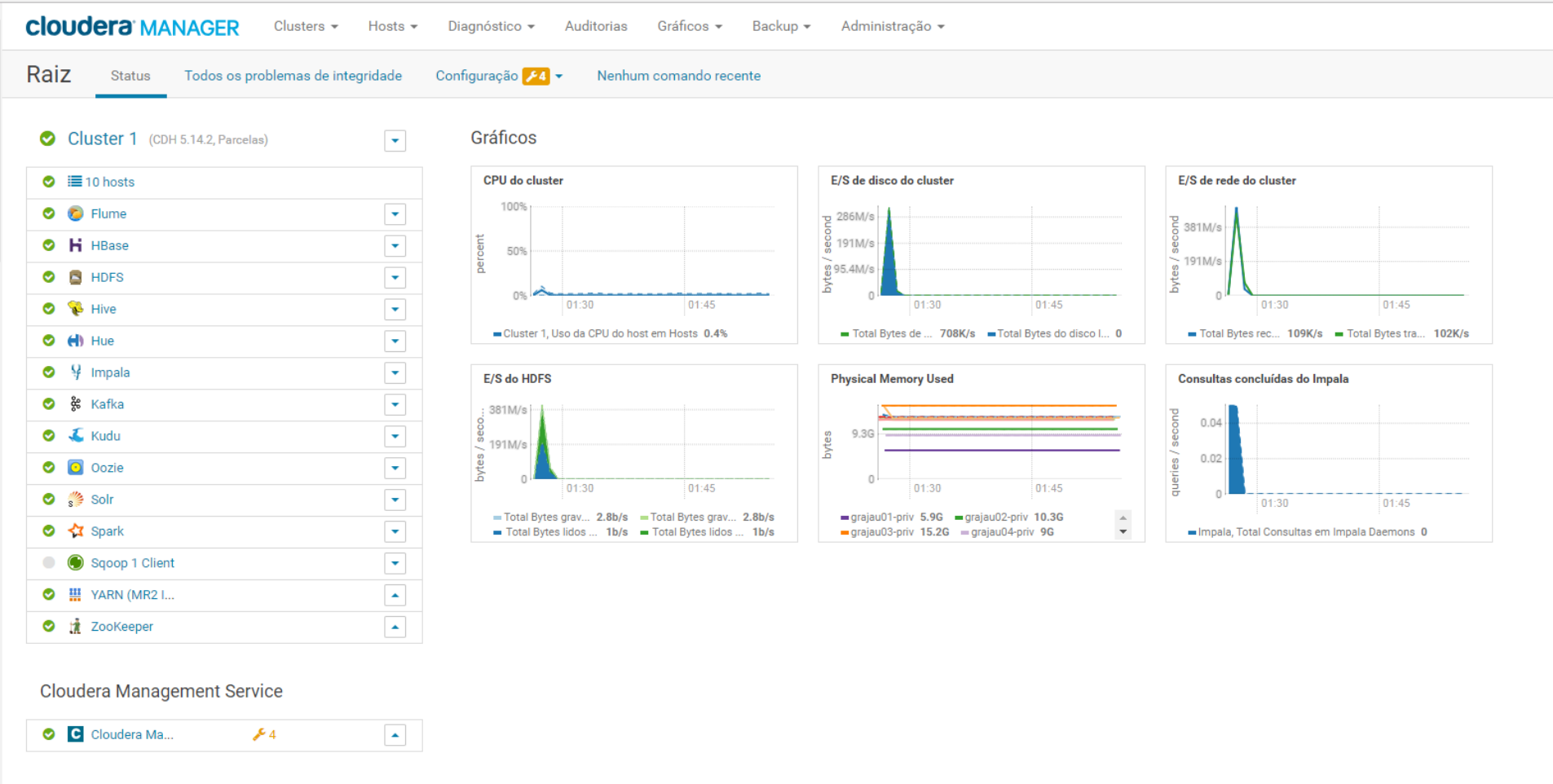
Ambiente Cloudera

10x Dell R730:
-2 processadores octa core;
-128GB de Memória RAM;
-06 Discos SAS de 4TB;
- Dual Port 10GB e 1GB.

Ambiente Cloudera



Ambiente Cloudera



Ambiente Cloudera

- Diferenças entre versão open source e subscrição:
    - Serviço de suporte 8x5 ou 24x7;
    - Atualizações periódicas;
    - Segurança e Integração com AD;
    - Gerenciar múltiplos clusters;

Ambiente Cloudera

- Aquisição:
  - PROFISCO no modelo de Pregão Eletrônico
  - Data do contrato: 01/02/2018

Item	Descrição	Unidade	Quantidade	Valor Unitário	Valor Total
1	Subscrição de software de ferramentas de BIG DATA por 24 meses - Cloudera Enterprise Data Hub Edition - Suporte 8x5	Nós	10	134.800,00	1.348.000,00
2	Serviço de desenvolvimento e implementação de soluções com repasse de conhecimento sobre ferramentas de BIG DATA	Horas	800	730,00	584.000,00
3	Serviço de treinamento sobre desenvolvimento e administração de ferramentas de BIG DATA - Turma de 10 pessoas - Carga horária de 120 horas	Turmas	4	115.700,00	462.800,00
	VALOR TOTAL: **R$ 2.394.800,00** (dois milhões trezentos e noventa e quatro mil e oitocentos reais)				

Caso#1: Faturamento NFCe

Problema

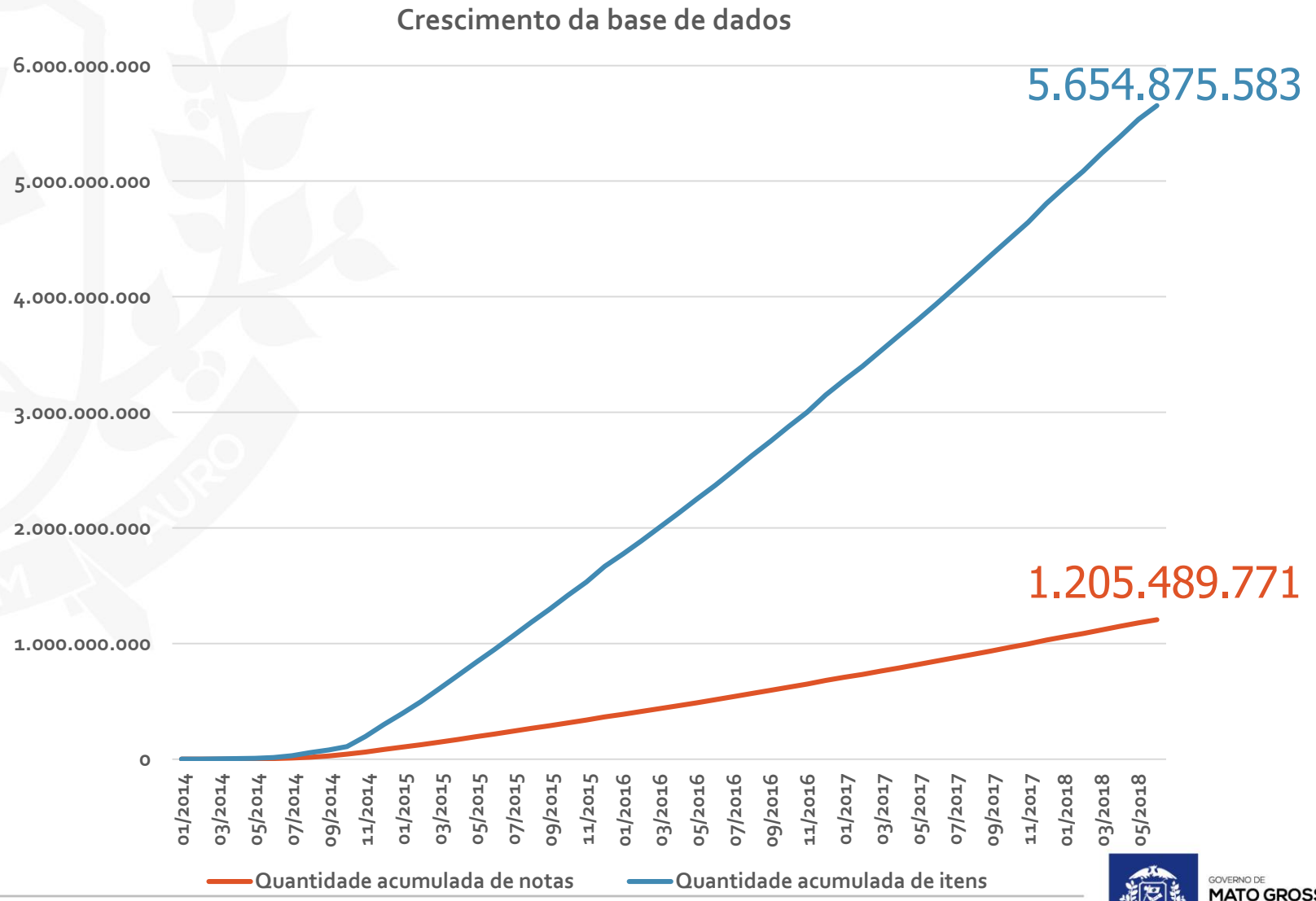
- Verificação de contribuintes enquadrados no SIMPLES:
  Calcular o faturamento mensal ou anual de todos os contribuintes, com a somatória de NFCE´s emitidas.

- A quantidade de informações na base é muito grande, tornando a consulta desenvolvida ineficaz.

Quantidade de dados

- NFCETB30_CONSUMIDOR_ELETRONICA
  - 14/11/2017 - 751.140.807
  - 16/04/2018 - 1.166.018.890
  - 26/06/2018 – 1.205.489.771
- NFCETB34_NFCE_PRODUTO_SERVICO
  - 14/11/2017 - 3.487.843.229
  - 16/04/2018 - 5.251.483.211
  - 26/06/2018 - 5.654.875.583
- Tamanho ocupado no cluster
  - NFCETB30_CONSUMIDOR_ELETRONICA
    - 316.05 G
  - NFCETB34_NFCE_PRODUTO_SERVICO
    - 231.95 G

Quantidade de dados



Arquitetura dos dados

Ambiente Autorização

- Modelo normalizado
- Armazenamento de XML

Ambiente Consulta

- Modelo denormalizado

Solução implementada

ORACLE EXADATA

Ambiente Consulta

- NFCETB30
- NFCETB34
- Dados Cadastrais

Ingestão dos novos dados

- Dados cadastrais
- Novas NFCe

DELL cloudera

Consolidação dos dados

Consulta executada

Consulta Oracle	• Não executa
Consulta Impala	• Tempo total de execução para um mês: 15 minutos

Ferramentas utilizadas

Sqoop	• Para fazer a carga dos dados
Hive	• Para fazer a ingestão inicial dos dados
Impala	• Para realizar a consulta após a conclusão da ingestão dos dados.

Estado atual e Planos futuros

Estado Atual

- Carga diária dos dados
- Resultado disponibilizado via reports
- Melhoria de performance

Planos Futuros

- Utilização dos conceitos em outras bases de dados grandes

Caso#2: Quantidade de NFCe emitida por contribuinte

Problema

- Contar quantas NFCes de cada tipo foram emitidas por cada um dos contribuintes.

- A quantidade de informações na base é muito grande, tornando a consulta desenvolvida ineficaz.

Consulta executada

Consulta Oracle

- Tempo total de execução para um mês: 40 minutos

Consulta Impala

- Tempo total de execução para um mês: 10 seg

Caso#3: Busca por acesso de IP específico

Problema

- Buscar usuário do acesso web que acessou por um determinado IP

- Nem todos os registro do Acesso Web no SCL possuíam o IP que era necessário

- Os dados do log do apache do servidor de produção possuem uma formatação específica

Solução implementada

- Pesquisar os logs dos servidores de produção
- Pesquisar quais requisições tiveram origem no IP buscado
- Buscar as funcionalidades acessadas pelo IP

```
1  192.168.5.3 - - [22/Sep/2016:08:17:55 -0400] "GET /" 200 422 "-" "-"
2  192.168.5.2 - - [22/Sep/2016:08:17:55 -0400] "GET /" 200 422 "-" "-"
3  177.**.**.152 - - [22/Sep/2016:08:17:56 -0400] "GET /imggd/imagens/miniatura/150x95/Array HTTP/1.1" 403 246 "http://www.sefaz.mt.gov.br/portal/index.php?link=1" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
   Chrome/53.0.2785.116 Safari/537.36"
4  191.***.***.101 - - [22/Sep/2016:08:17:56 -0400] "GET /arrecadacao/darlivre/menudarlivre HTTP/1.1" 200 6515 "-" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.116 Safari/537.36"
5  201.**.***.130 - - [22/Sep/2016:08:17:56 -0400] "GET /sid/consulta/infocadastral/consultar/publica HTTP/1.1" 200 9035 "http://www.sintegra.gov.br/new_bv.html" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
   Chrome/53.0.2785.116 Safari/537.36"
6  191.***.***.101 - - [22/Sep/2016:08:17:56 -0400] "GET /estilos/SefazEstilos.css HTTP/1.1" 304 - "https://www.sefaz.mt.gov.br/arrecadacao/darlivre/menudarlivre" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko)
   Chrome/53.0.2785.116 Safari/537.36"
```

Ferramentas utilizadas

HDFS	• Armazenamento inicial dos arquivos de logs
HIVE	• Ingestão dos dados • Transformação de vários arquivos de log em uma única tabela HIVE
Impala	• Consulta dos dados
Jupyter Notebook	• Analise dos dados

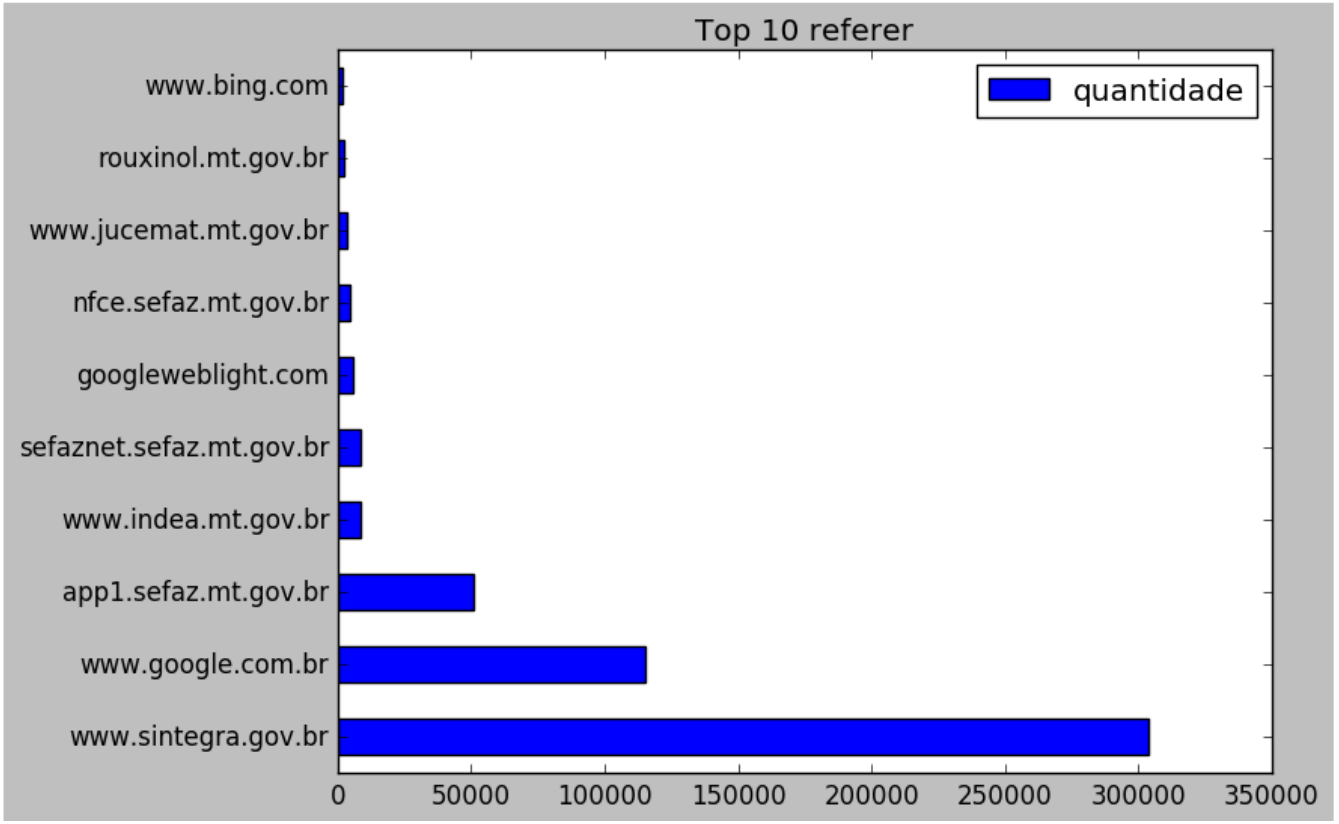
Quantidade de dados

- Arquivos originais:
  - 96 arquivos
  - 23,4Gb
- LOGTB01_ACCESS_LOG
  - 116.283.297
- Tamanho ocupado no cluster
  - LOGTB01_ACCESS_LOG
    - 5.2Gb

Top 10 Referer

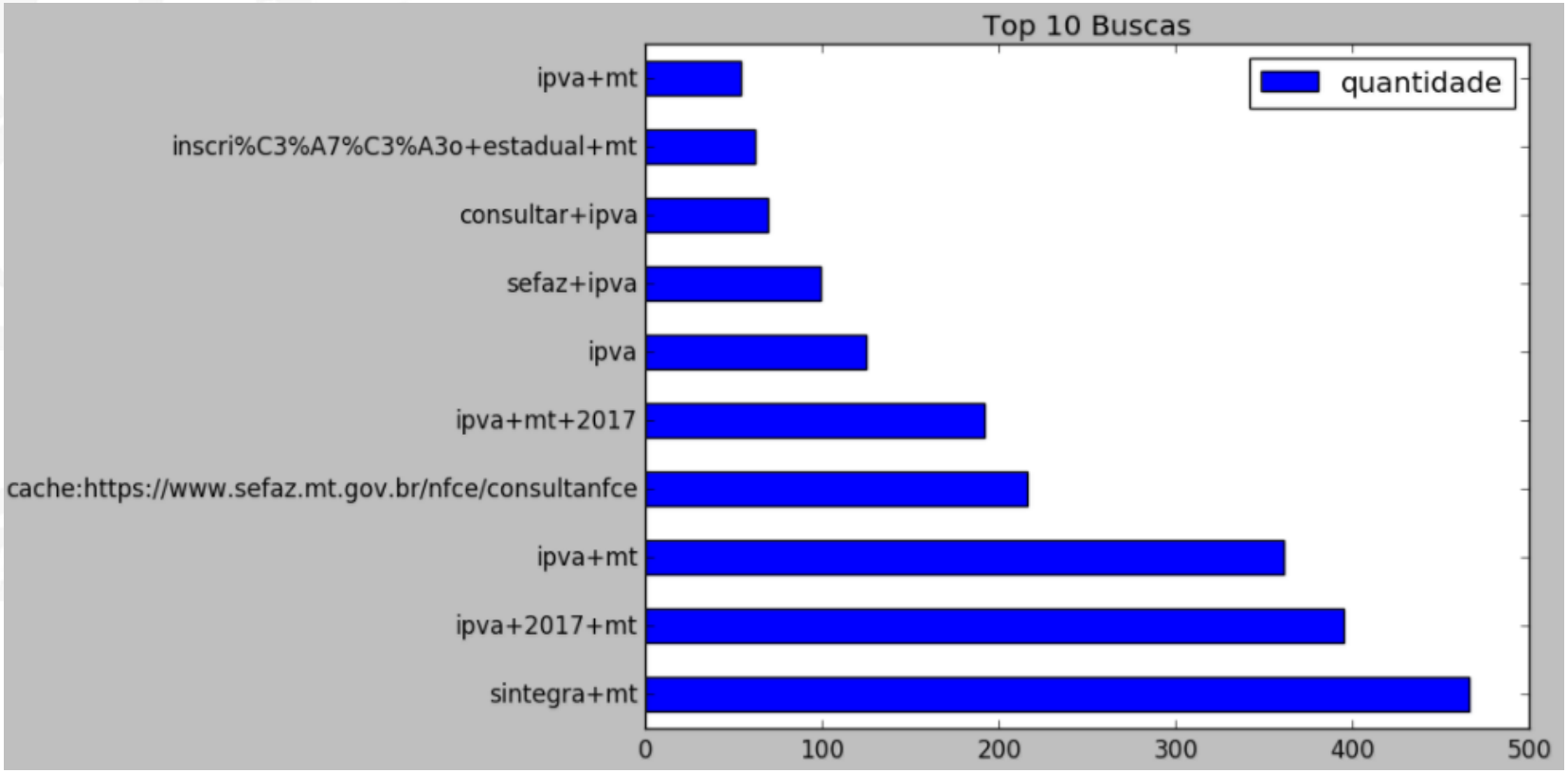

```
top10 = df_top_referers[(df_top_referers.referer_host != 'www.sefaz.mt.gov.br') & (df_top_referers.referer_host != '')].head(10)
ax = top10.plot(kind='barh', title='Top 10 referer')
#ax.axis('off')

ax.set_yticklabels([i for i in top10.referer_host])
rects = ax.patches
```



Top 10 Buscas

Busca por IP específico



Top 10 Buscas

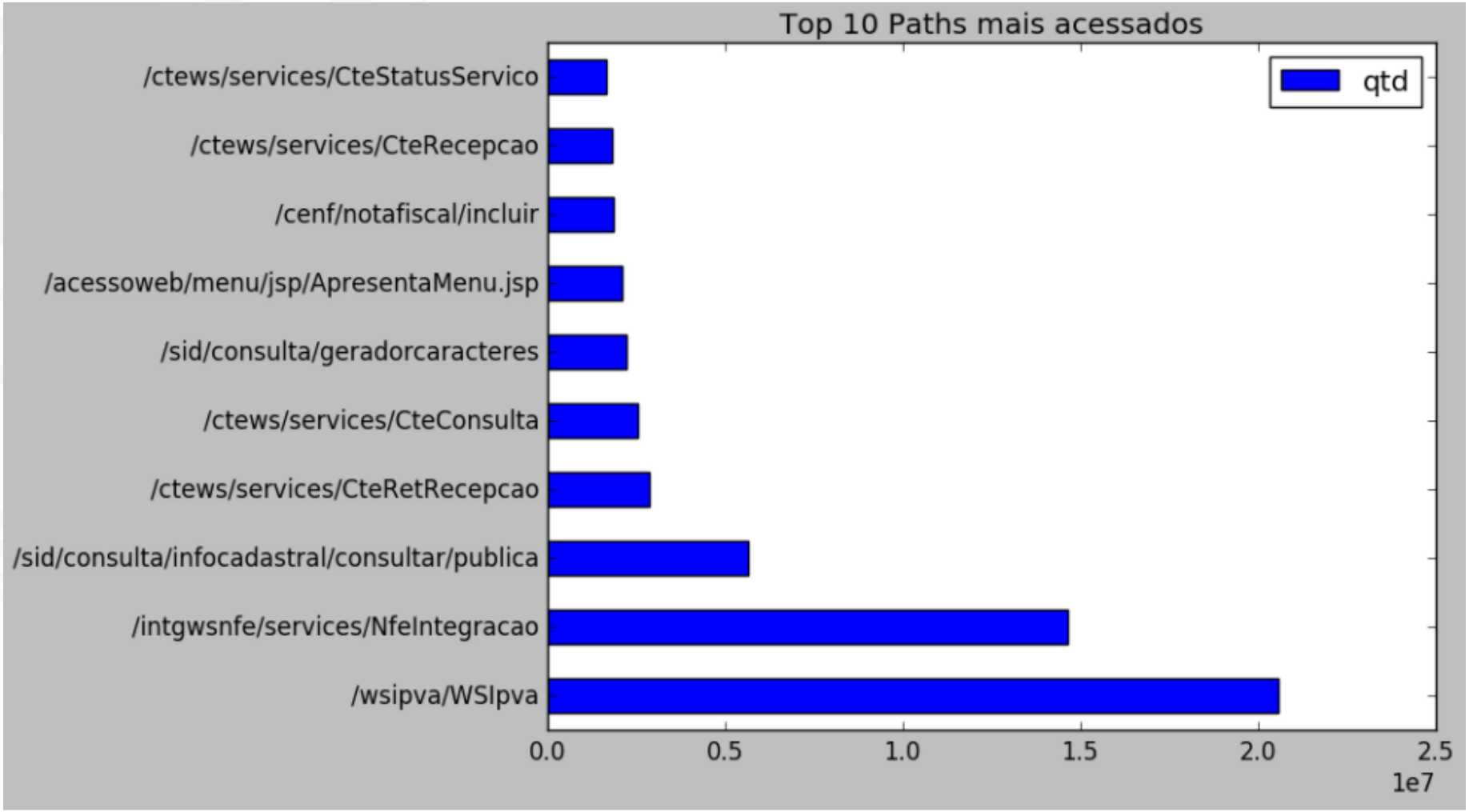
Busca por IP específico

```
df_searchs.head(10)
```

	referer_host	query	quantidade
0	www.google.com.br	sintegra+mt	466
1	www.google.com.br	ipva+2017+mt	395
2	www.google.com.br	ipva+mt	361
3	webcache.googleusercontent.com	cache:https://www.sefaz.mt.gov.br/nfce/consult...	216
4	www.google.com.br	ipva+mt+2017	192
5	www.google.com.br	ipva	125
6	www.google.com.br	sefaz+ipva	99
7	www.google.com.br	consultar+ipva	69
8	www.google.com.br	inscri%C3%A7%C3%A3o+estadual+mt	62
9	www.bing.com	ipva+mt	54

Top 10 Paths mais acessados

Busca por IP específico



Top 10 Paths mais acessados


```
df_mais_acessados.head(10)
```

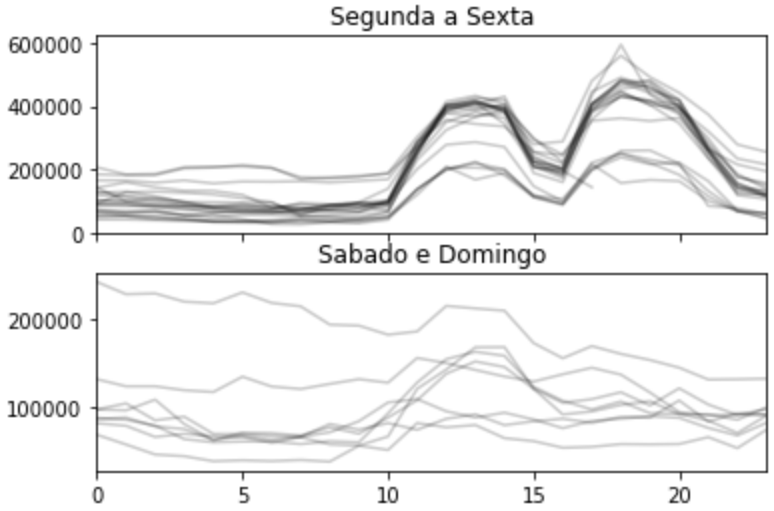
	path1	qtd
0	/wsipva/WSIpva	20560252
1	/intgwsnfe/services/NfeIntegracao	14652998
2	/sid/consulta/infocadastral/consultar/publica	5672614
3	/ctews/services/CteRetRecepcao	2864434
4	/ctews/services/CteConsulta	2567251
5	/sid/consulta/geradorcaracteres	2243567
6	/acessoweb/menu/jsp/ApresentaMenu.jsp	2092318
7	/cenf/notafiscal/incluir	1864221
8	/ctews/services/CteRecepcao	1815906
9	/ctews/services/CteStatusServico	1662677

Padrões de acesso aos sistemas fazendários

Busca por IP específico

```
In [14]: f, ax = plt.subplots(2, sharex=True)
         plt.xlim((0,23))
         for day in dias:
             x, y, weekday = get_serie(day)
             if weekday[0] < 5:
                 ax[0].plot(x, y, alpha=0.2, label=day, c='black')
             else:
                 ax[1].plot(x, y, alpha=0.2, label=day, c='black')
         ax[0].set_title('Segunda a Sexta')
         ax[1].set_title('Sabado e Domingo')
```

```
Out[14]: <matplotlib.text.Text at 0xb3072e8>
```



Estado atual e Planos futuros

Estado Atual

- Carga inicial dos logs feita
- Buscas pelos IPs que foram específicados
- Algumas outras análises

Planos Futuros

- Criar um processo para salvar os logs no cluster
- Continuar o processo de análise e monitoramento dos logs

Perguntas?