Armazenando e Processando Documentos Fiscais Eletrônicos com o Hadoop, Spark e NiFi

Fábio Andreatta – SEFAZ/MS

Agenda

- O que é o Hadoop?
- O Projeto
- Trabalhos Futuros

O Que é o Hadoop?

O Que é o Hadoop?

2003
- O Google publica o artigo sobre Google File System

2004
- O Artigo sobre MapReduce é publicado

2008
- Começa o desenvolvimento do YARN
- O motor de pesquisa do Yahoo!

2010
- HBase e Hive: dois dos principais banco de dados do ecosistema
- Yahoo! com um cluster com mais de 4000 nós

2011
- Facebook, LinkedIn, eBay e IBM contribuem com mais de 20 mil linhas de código

O Que é o Hadoop?

- Hoje, o Hadoop é uma coleção de programas de código aberto que facilitam o uso de um cluster de computadores para resolver problemas envolvendo grande quantidade de dados.

Processamento

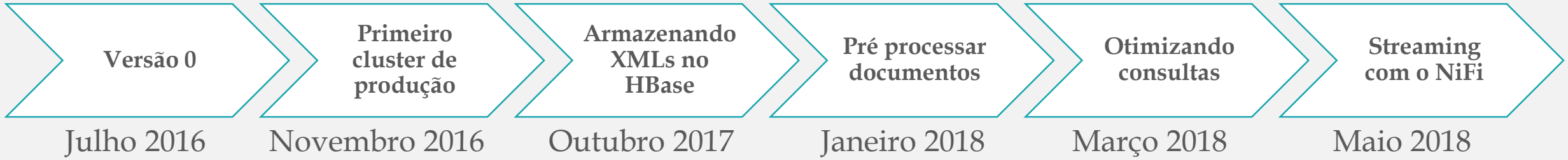
Armazenamento

Stream, ETL e Sincronização

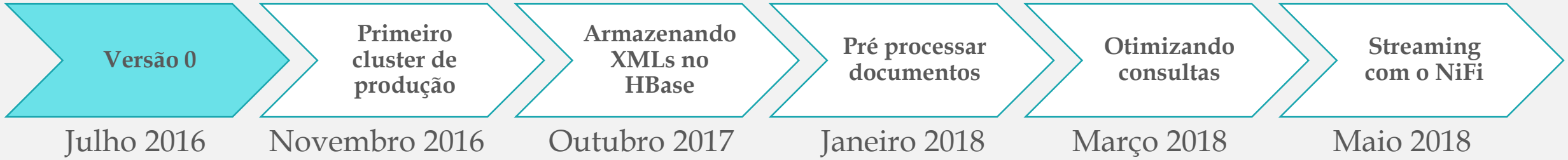
O Projeto

Armazenar e consultar dados de Documentos Fiscais Eletrônicos (DF-e) no Hadoop

As Etapas do Projeto

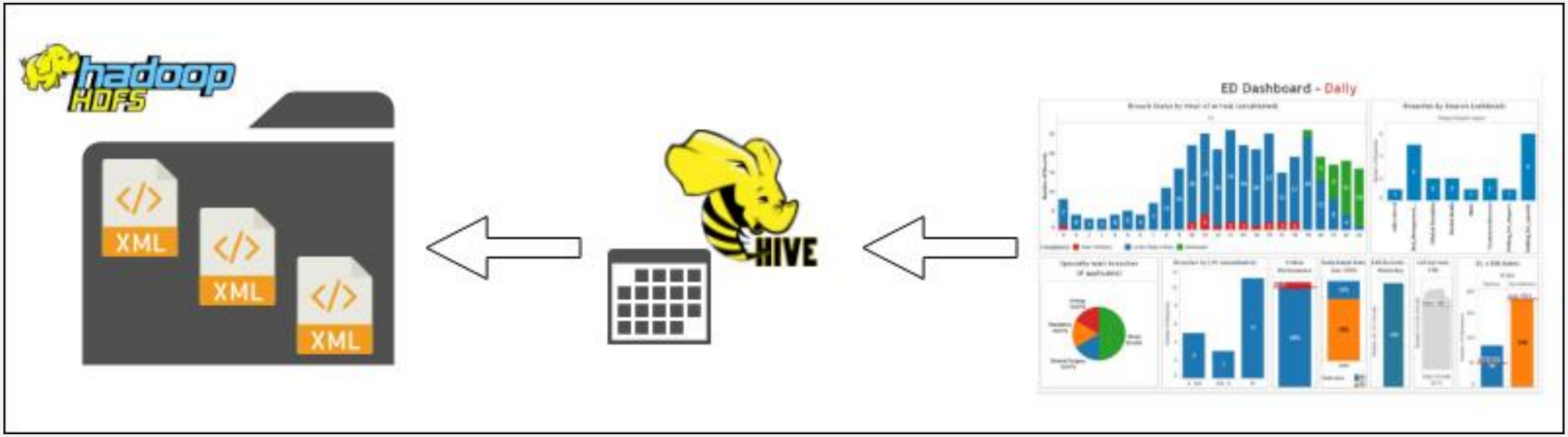


As Etapas do Projeto



Versão 0

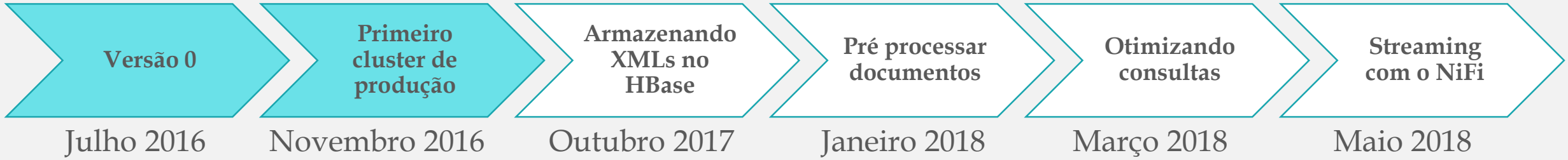
- Armazenar arquivos diretamente no HDFS
- Utilizamos o Hive com a biblioteca SerDe para efetuar a leitura dos arquivos XML (External Table)



Versão 0

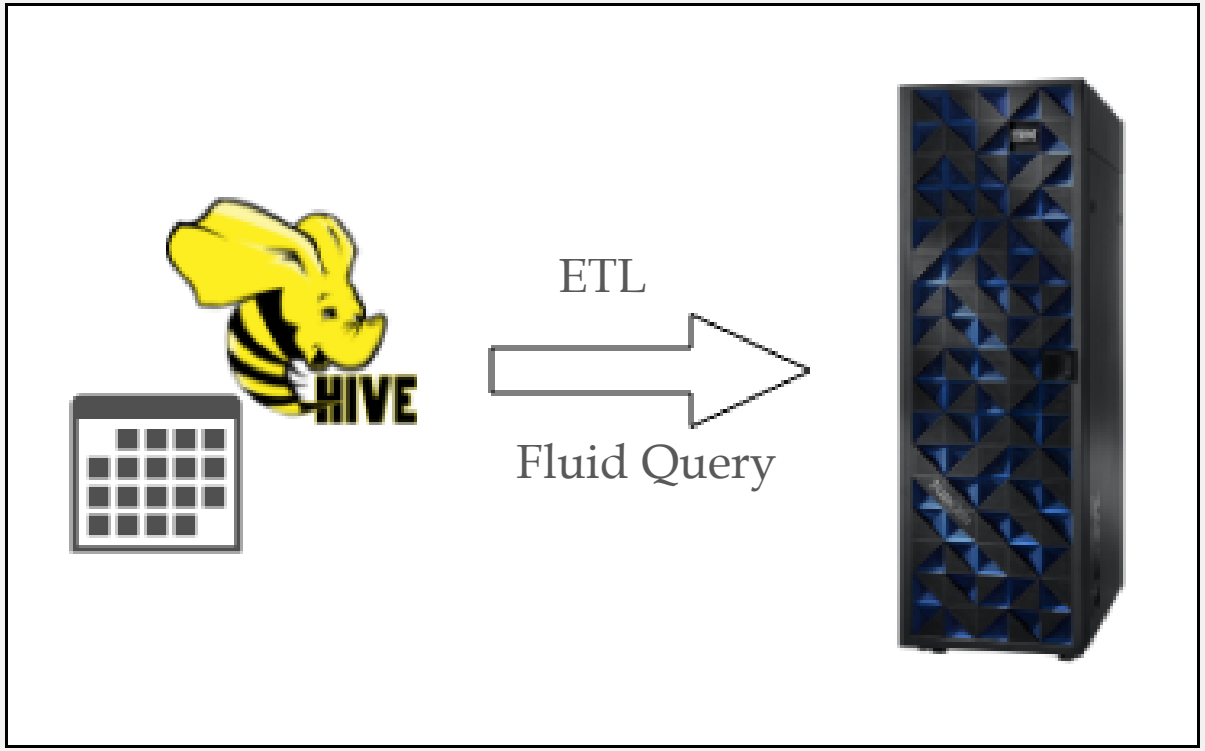
- Muitos arquivos pequenos armazenados no HDFS
- Consultas ficaram muito lentas: o hive precisa interpretar todos os xml em cada consulta
- Precisaríamos de uma infraestrutura com alto poder de processamento para resolver consultas em tempo adequado

As Etapas do Projeto

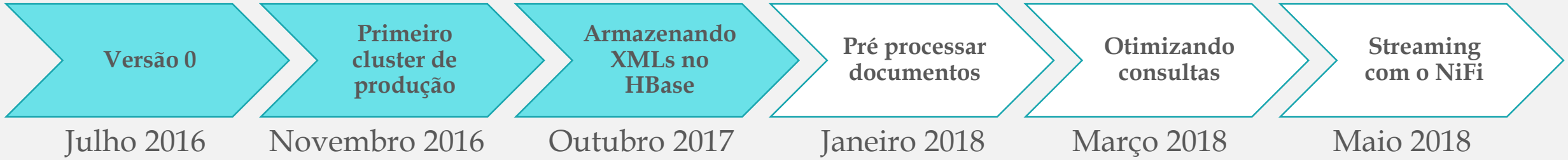


Primeiro cluster de Produção

- Reciclamos três servidores
  - 4TB de Armazenamento;
  - 60GB de Memória;
  - 24 Núcleos de Processamento.
- Integrar o Hadoop com o IBM Netezza (Plataforma Analítica)
  - ETL (IBM DataStage)
  - FluidQuery
- Identificação dos itens de NFC-e para calcular o preço médio ponderado ao consumidor final (Pauta)



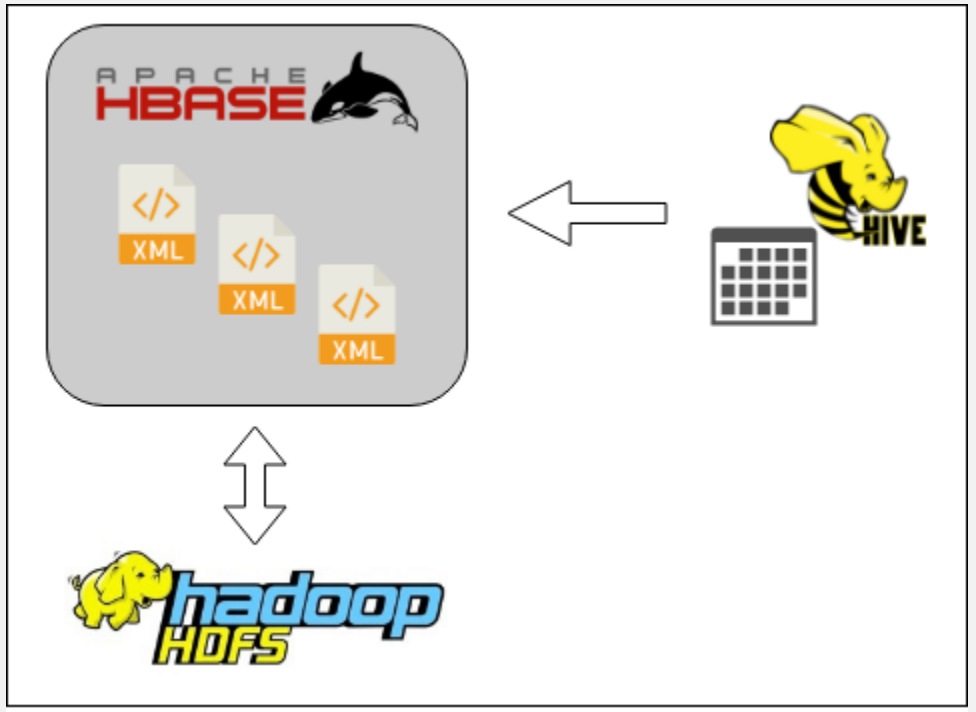
As Etapas do Projeto



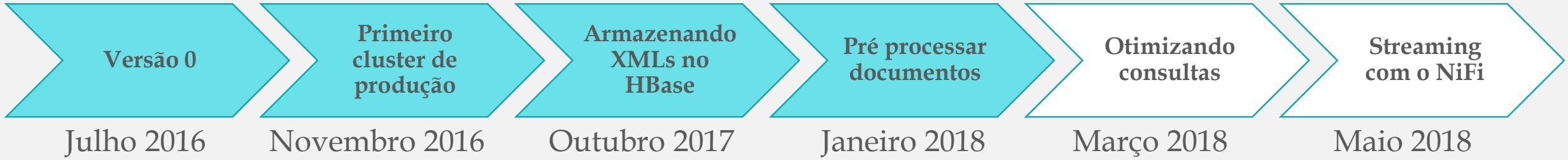
Armazenando XMLs no HBase

- Tiramos os dados do HDFS e passamos a salvá-los no HBase
- Mapeamos o Hive para ler os dados do HBase (External Table)

- O Hive continua fazendo full scan na base inteira para qualquer consulta
- O tempo de execução das consultas melhorou, mas ainda não podiam ser usadas diretamente.

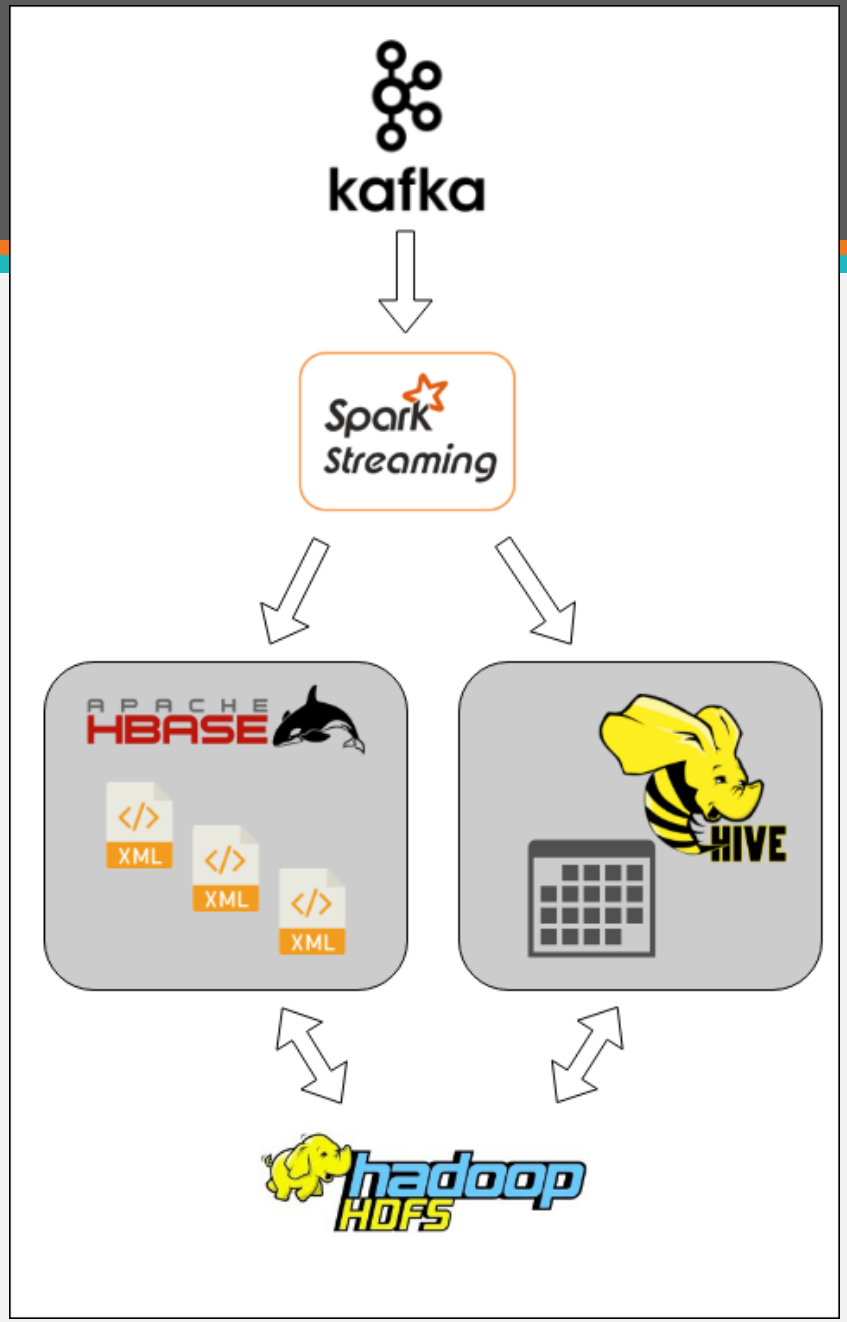


As Etapas do Projeto



Armazenando XMLs no HBase

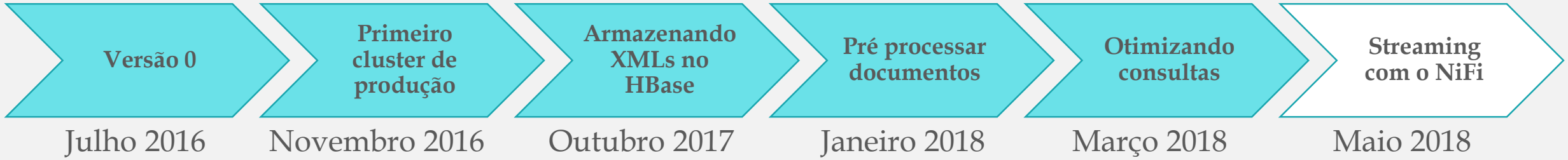
- Criamos tabelas gerenciadas pelo Hive (Managed Tables) no formato ORC
- Passamos a pré processar os documentos no momento em que chegam no Hadoop (logo após a autorização)
- O Hive não precisou mais fazer full scans
- O formato ORC pode ser compactado
- O tempo de algumas consultas ficou aceitável



O Spark

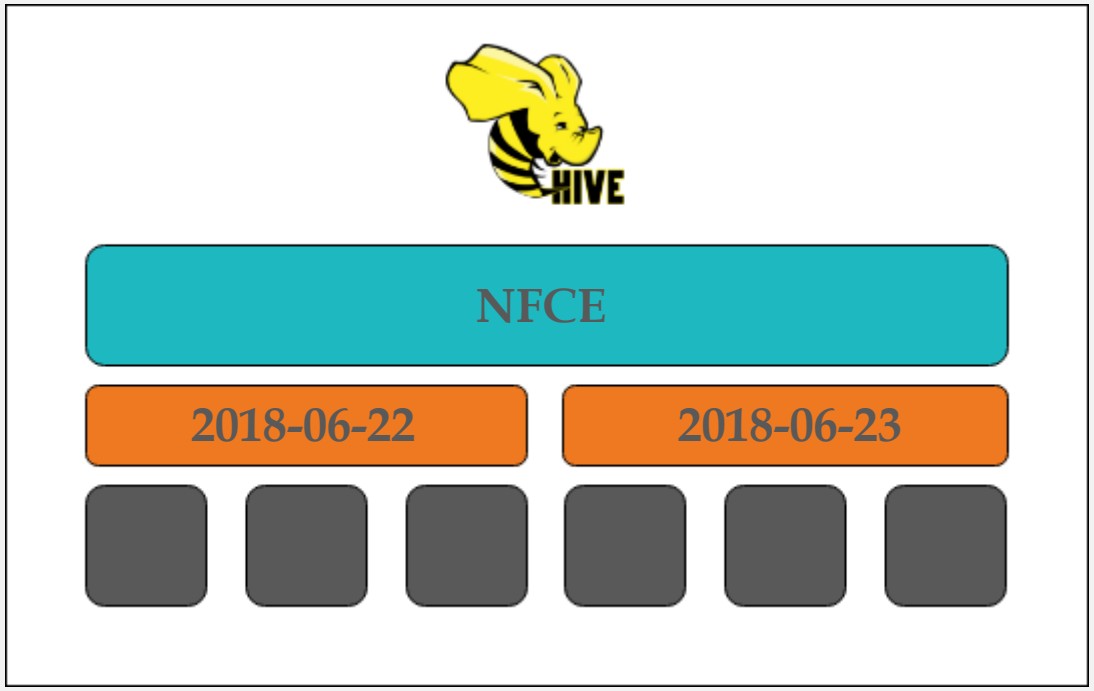
- Processamento em memória (Muito Rápido)
- Biblioteca de Machine Learning com os principais algorítimos necessários (Spark MLLib)
- Consultas SQL (Spark SQL)

As Etapas do Projeto



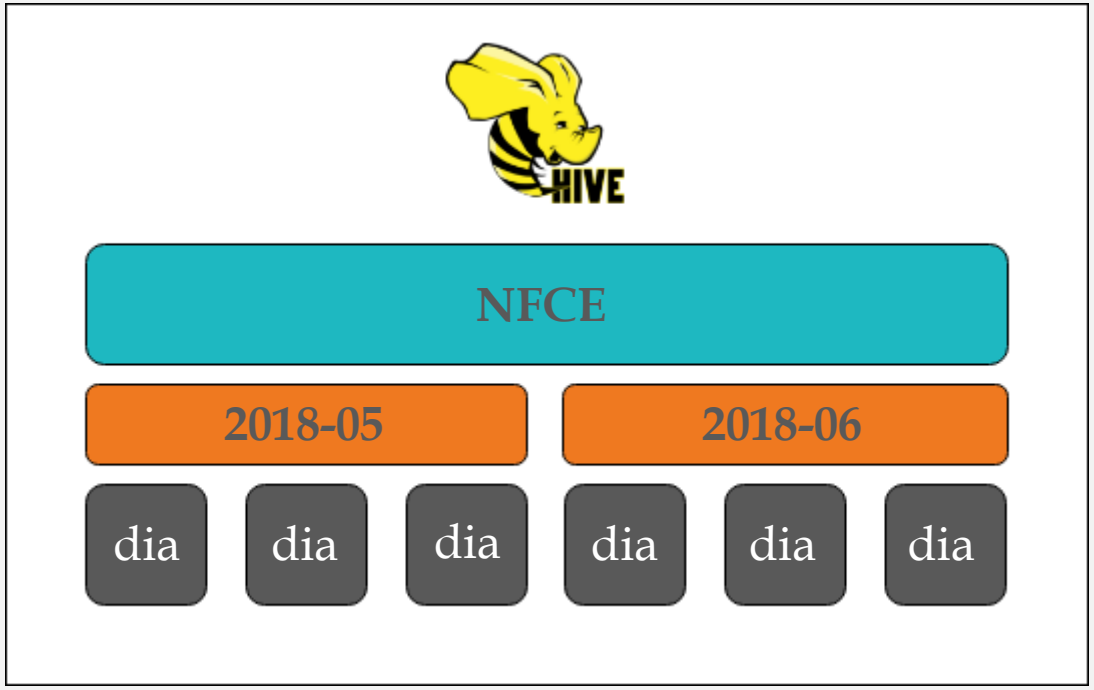
Otimizando as consultas

- Tentativa #1
- As tabelas foram particionadas por dia:
  - As consultas ficaram com um tempo de resposta muito melhor
  - Mas os arquivos de dados ficaram muito pequenos



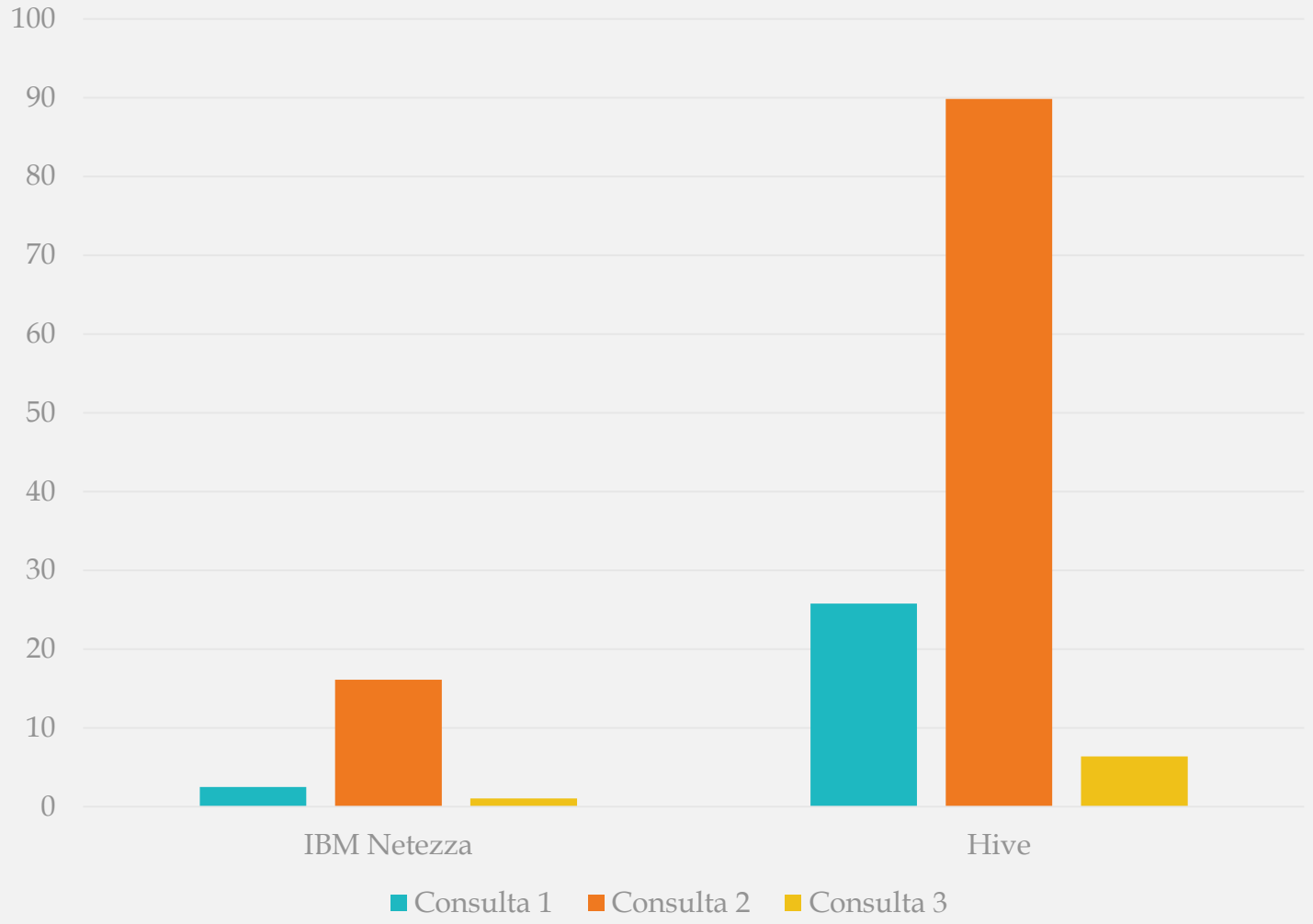
Otimizando as consultas

- Tentativa #2
- As tabelas foram particionadas por mês e os arquivos de dados ordenados por dia:
  - O tempo de resposta das consultas melhoraram ainda mais
  - Os arquivos de dados ficaram com tamanhos entre 100MB e 300MB



Resultado das otimizações

- **Consulta #1**
  - Contagem do número de NFCe por dia nos últimos 90 dias
- **Consulta #2**
  - Separar os itens de NFCe que representam 65% do faturamento nos últimos 90 dias
- **Consulta #3**
  - Contagem do número de Nfes canceladas nas últimas 24h

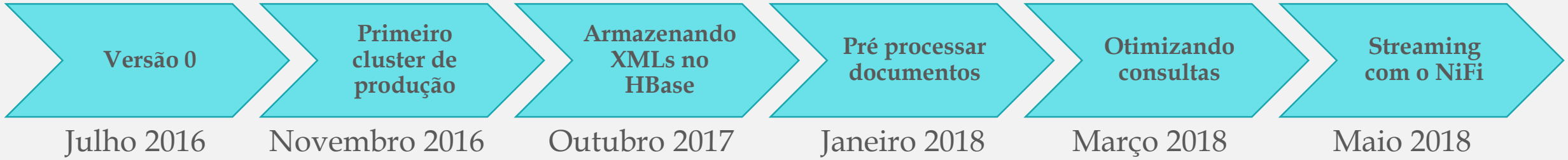


Resultado das otimizações

- 32TB para armazenamento
- 40 CPU Cores
- 32 FPGA Cores

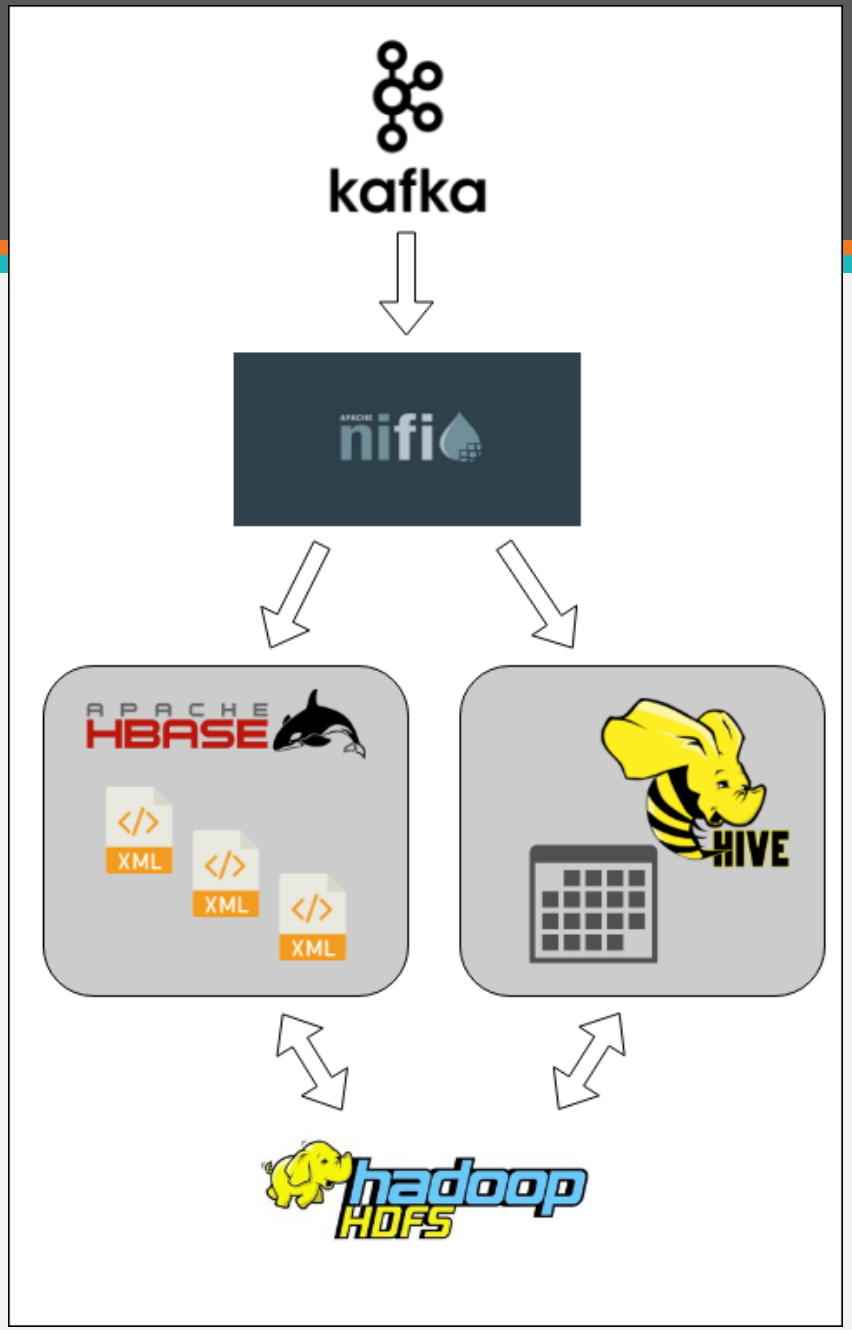
- 22TB para armazenamento
- 16 vCores

As Etapas do Projeto

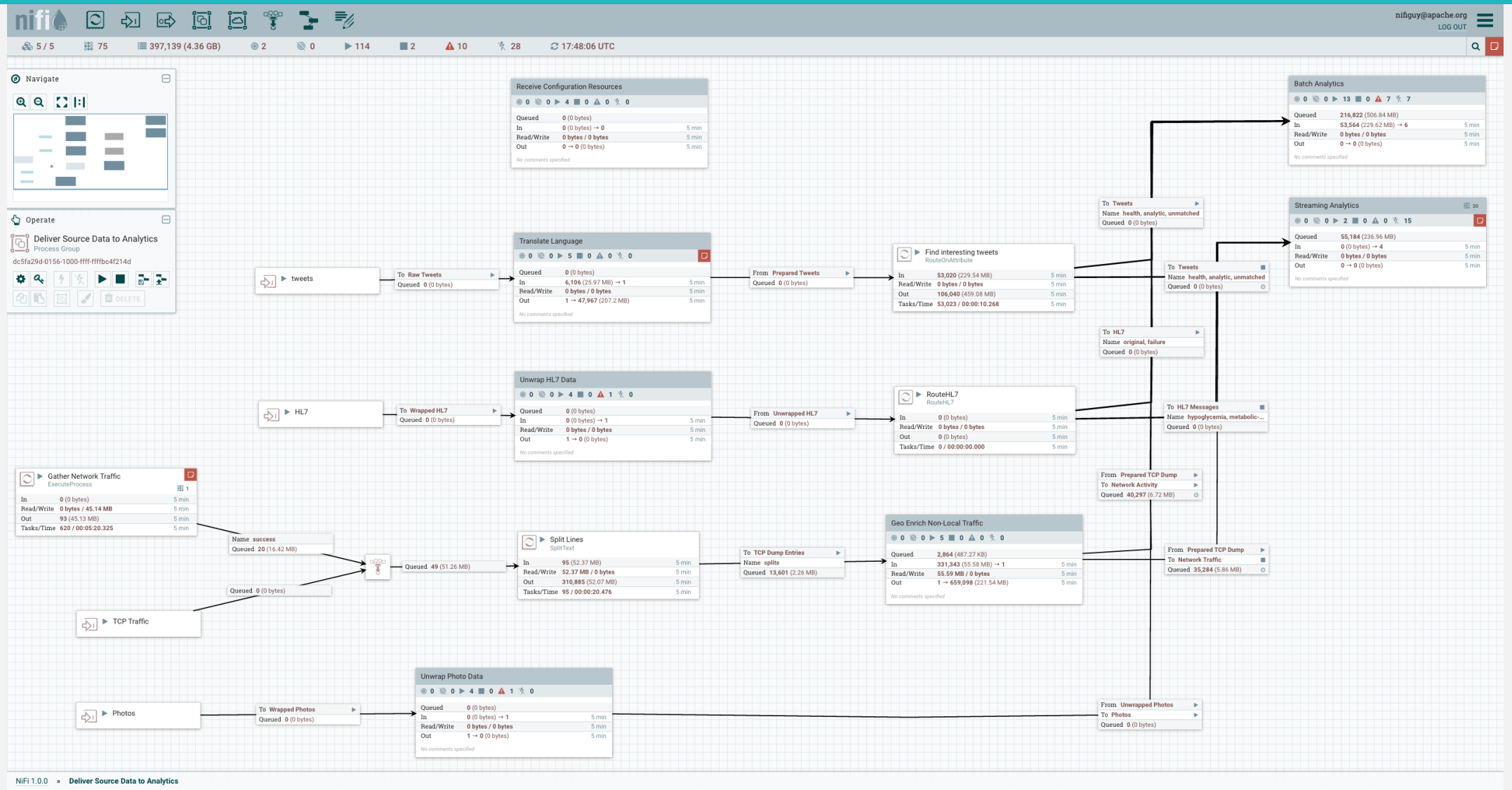


Utilizando o NiFi para Pré Processar os XMLs

- Menos manutenção em código
- Monitoramento do fluxo de dados
- Dados disponíveis no ambiente analítico em tempo **quase real** (aproximadamente 60 minutos)



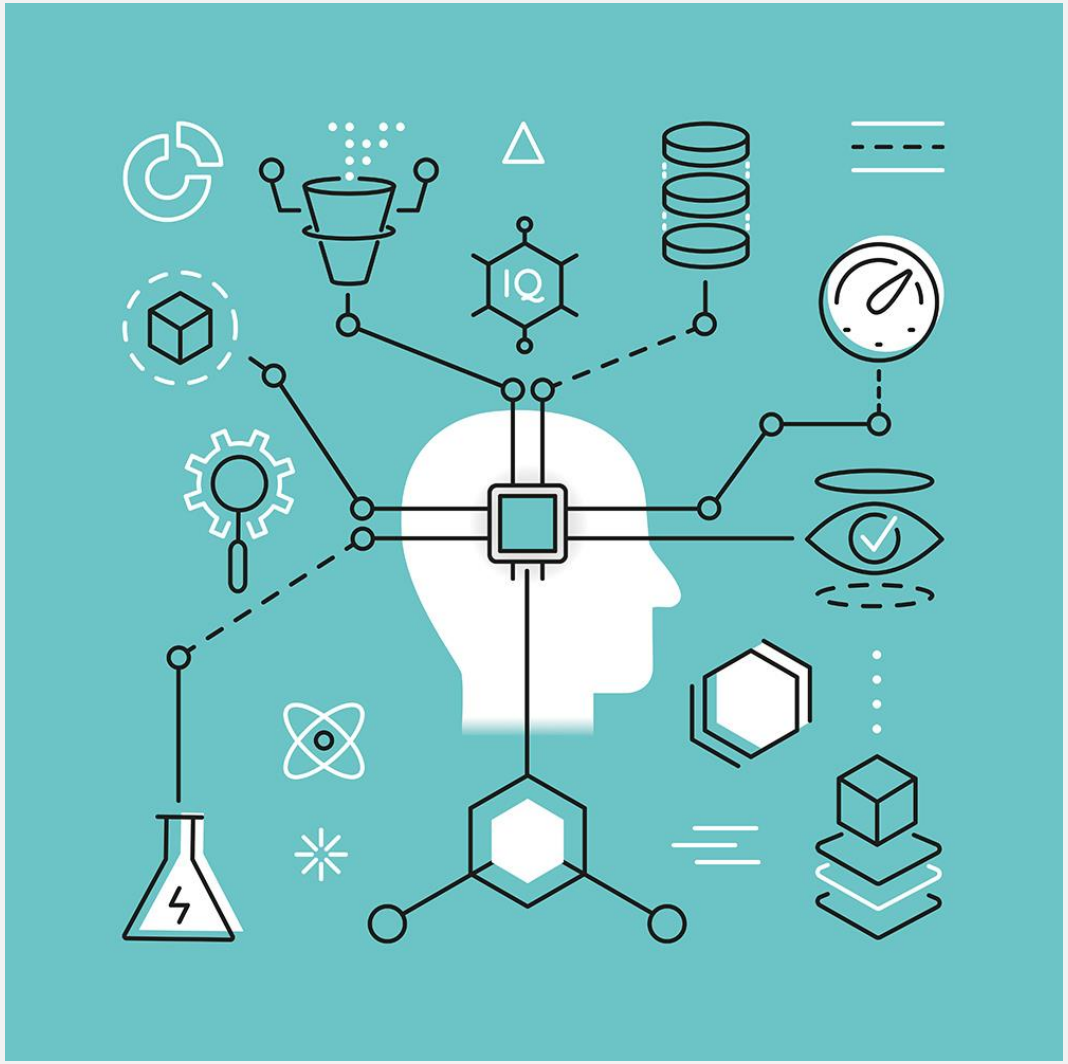
O NiFi


nifiguy@apache.org
LOG OUT
5 / 5
75
397,139 (4.36 GB)
2
0
114
2
10
28
17:48:06 UTC
Navigate
Operate
Deliver Source Data to Analytics
Process Group
dc5fa29d-0156-1000-ffff-ffffbc4f214d
DELETE
Receive Configuration Resources
Batch Analytics
Streaming Analytics
tweets
To Raw Tweets
Translate Language
From Prepared Tweets
Find interesting tweets
RouteOnAttribute
To Tweets
Name health, analytic, unmatched
HL7
To Wrapped HL7
Unwrap HL7 Data
From Unwrapped HL7
RouteHL7
To HL7
Name original, failure
To HL7 Messages
Name hypoglycemia, metabolic-...
Gather Network Traffic
ExecuteProcess
Name success
Split Lines
SplitText
To TCP Dump Entries
Name splits
Geo Enrich Non-Local Traffic
From Prepared TCP Dump
To Network Activity
To Network Traffic
TCP Traffic
Photos
To Wrapped Photos
Unwrap Photo Data
From Unwrapped Photos
To Photos
NiFi 1.0.0 » Deliver Source Data to Analytics

Trabalhos Futuros

Trabalhos Futuros

- Utilizar o HBase para armazenar arquivos binários
  - Tirar BLOBs dos bancos de dados transacionais SQL
- Aplicar algoritmos de Machine Learning para apoiar a fiscalização



Muito obrigado! ☺