

DAS - Documento de Arquitetura de Solução

Carga de Dados Financeiros

Autor: Fernando Alves da Costa



Sumário

Resumo	<i>3</i>
1 – Escopo do Projeto	4
2 – Desenho de Arquitetura	4
3 – Hardware e Software utilizados	5
4 – Dados	6
5 - Detalhes da solução	6
6 – Conclusão	13



Resumo

O presente documento apresenta uma descrição macro da implantação projeto Carga de Dados Financeiros. O objetivo desse documento é detalhar as ferramentas utilizadas.



1 – Escopo do Projeto

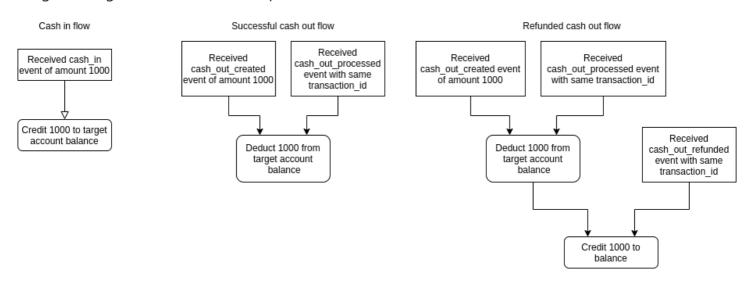
Dado um conjunto de tópicos do Kafka que descreve vários estados de transações financeiras, crie um aplicativo que processe esses eventos quase em tempo real e gere um saldo consolidado para um determinado par account_id e datetime solicitado.

- Os saldos devem:
 - o incluem opções para filtrar por account_id, datetime e contraparte quando solicitado
 - mostrar todo o histórico de transações até o momento selecionado junto com o valor disponível na conta naquele momento.
- A aplicação deve ser preferencialmente escrita em Scala, Python ou Elixir.
- O reprocessamento dos eventos não deve levar a um estado que diverge da execução original.

O conjunto de dados é composto por 4 arquivos JSON, cada um dos quais deve ser processado como um tópico Kafka separado, como está. Etapas de processamento adicionais podem ser executadas posteriormente. Os estados de saque são representados como:

- criado: criado, mas ainda a ser processado, nenhum dinheiro deve sair da conta
- processado: a transação foi processada com sucesso, portanto, o saldo total deve ser deduzido do valor da transação correspondente
- reembolsado: a transação não foi concluída, o valor original deve retornar para a conta

A figura a seguir ilustra os cenários possíveis:



2 – Desenho de Arquitetura

De acordo com a necessidade da área de negócio, foi definido uma arquitetura que atendesse por completo ao projeto.



Abaixo está o desenho macro da solução.

Stone

Desenho de Arquitetura

Fonte de Dados

Integração

Processamento

StreamSets

StreamSets

Visualização dos Dados

Visualização dos Dados

3 – Hardware e Software utilizados

Desktop

Ferramenta/Hardware	Função
Hardware Macbook i7, 1tb de Disco, 16gb de	Hospedar o software de virtualização e as máquinas
Memória	virtuais
Parallels Desktop	Software de virtualização de servidores
Máquina Virtual	Servidores Virtuais, compartilhando o mesmo
	hardware
CentOs	Sistema Operacional Linux, usando nas máquinas
	virtuais
Postgres	Banco de Dados relacional
StreamSets	Coletor de Dados, usado para fazer ingestão de dados
Kafka	plataforma de streaming de eventos distribuídos,
	usado em análise de streaming, integração de dados,
	etc

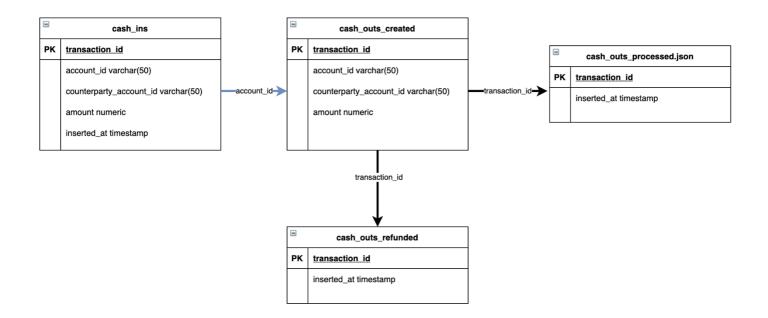


PowerBI Desktop	Ferramenta para visualização de dados, criação de
	Dashboard e relatórios.
Python	Linguagem de programação OpenSource
Apache Spark	Apache Spark é um mecanismo de análise unificado
	para processamento de dados em grande escala.

4 - Dados

• Foram disponibilizados pelo usuário quatro arquivos no formato json.

```
cash_outs_created.json
cash_outs_refunded.json
cash_outs_processed.json
cash_ins.json
```



5 - Detalhes da solução

1) Postgres – Criando um banco de dados para armazenar os quatros arquivos json, foi criada uma tabela para cada arquivo.

```
CREATE TABLE cash_ins (
transaction_id varchar(50) primary key,
account_id varchar(50),
counterparty_account_id varchar(50),
amount numeric,
inserted_at timestamp);

CREATE TABLE cash_outs_created (
transaction_id varchar(50) primary key,
```



```
account_id varchar(50),
counterparty_account_id varchar(50),
amount numeric,
inserted_at timestamp)

CREATE TABLE cash_outs_processed (
transaction_id varchar(50) primary key,
inserted_at timestamp)

CREATE TABLE cash_outs_refunded (
transaction_id varchar(50) primary key,
inserted_at timestamp)
```

2) Kafka – Criado quatro tópicos no Kafka, cada tópico para armazenar seu arquivo correspondente.

```
/opt/kafka/bin/kafka-topics.sh --create --topic cash_ins --zookeeper localhost:2181 --replication-factor 1 --partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
--partitions 1
```

- 3) StreamSets Criado quatro coletores, um para cada arquivo json. O coletor irá monitorar o arquivo json, caso um novo registro seja inserido, o mesmo será enviado para o tópico Kafka.
- 4) Python Criado um programa que fica monitorando os tópicos no Kafka, caso chegue um novo registro, o mesmo é inserido no banco de dados em tempo real. Para a execução do programa, passar o nome do tópico por parâmetro.

```
# Programa para carga de topico no kafka
# Autor: Fernando Alves da Costa
# Data Criação: 14/01/2021
# Bibliotecas utilizadas
from kafka import KafkaConsumer
from json import loads
import psycopg2
import ast
import sys
import datetime
# Tratando os parametros de entrada
param1=[]
param1=str(' '.join(sys.argv[1:])).split()
if param1 == ":
 print("Favor, passar o parametro com o nome do tópico")
 sys.exit()
else:
 print("Parametro encontrado")
```



```
# Parametros de banco de dados
host = "postgres02"
dbname = "dw"
user = "dw"
password = "Dw01"
port = "5432"
# inicio do loop, o programa foi feito para passar varios tópicos como entrada
for param in param1:
 print("Processando o topico: ",param)
 # Conexao com o banco de dados
 conn_string = "host={0} user={1} dbname={2} password={3} port={4}".format(host, user, dbname, password,
 conn = psycopg2.connect(conn_string)
    print("Sucesso ao conectar no banco de dados")
    print("Erro ao conectar no banco de dados")
    sys.exit()
 # Conexao com o kafka com metódo consumer
 # Para consumir o tópico desde o inicio, alterar o parametro auto offset reset='earliest'
 # Para nao deixar o programa rodando sem parar, incluir o parametro consumer_timeout_ms = 15000
 consumer = KafkaConsumer(param, bootstrap servers='kafka01:9092', consumer timeout ms = 15000,
auto_offset_reset='latest', enable_auto_commit=True)
    print("Sucesso ao conectar no kafka")
 except:
    print("Erro ao conectar no kafka")
 # Verifica qual o tópico, para definir tabela no postgres
 def verifica layout arquivo():
    global transaction_id_
    global account_id_
    global counterparty_account_id_
    global amount_
    global inserted_at_
    global postgres_insert_query
    global record_to_insert
    # Preenchendo as variaveis para carga no postgres
    if param == 'cash ins':
      transaction_id_=mydata.get('transaction_id')
      account_id_=mydata.get('account_id')
      counterparty_account_id_=mydata.get('counterparty_account_id')
      amount =mydata.get('amount')
      inserted at =datetime.datetime.fromtimestamp(mydata.get('inserted at') / 1e3)
      postgres insert_query = """ insert into cash_ins (transaction_id, account_id, counterparty_account_id,
amount, inserted_at) values (%s, %s, %s, %s, %s, %s)"""
      record_to_insert = (transaction_id_, account_id_, counterparty_account_id_, amount_, inserted_at_)
    elif param == 'cash outs created':
```



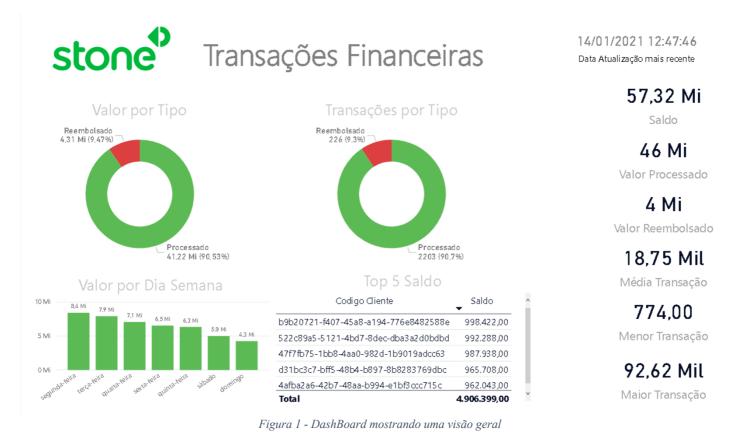
```
transaction_id_=mydata.get('transaction_id')
      account_id_=mydata.get('account_id')
      counterparty account id =mydata.get('counterparty account id')
      amount_=mydata.get('amount')
      inserted_at_=datetime.datetime.fromtimestamp(mydata.get('inserted_at') / 1e3)
      postgres insert query = """ insert into cash outs created (transaction id, account id,
counterparty account id, amount, inserted at) values (%s, %s, %s, %s, %s, %s)"""
      record_to_insert = (transaction_id_, account_id_, counterparty_account_id_, amount_, inserted_at_)
    elif param == 'cash_outs_refunded':
      transaction id =mydata.get('transaction id')
      inserted at =datetime.datetime.fromtimestamp(mydata.get('inserted at') / 1e3)
      postgres_insert_query = """ insert into cash_outs_refunded (transaction_id, inserted_at) values (%s, %s)"""
      record to insert = (transaction id , inserted at )
    elif param == 'cash outs processed':
      transaction_id_=mydata.get('transaction_id')
      inserted_at_=datetime.datetime.fromtimestamp(mydata.get('inserted_at') / 1e3)
      postgres insert_query = """ insert into cash_outs_processed (transaction_id, inserted_at) values (%s, %s)"""
      record_to_insert = (transaction_id_, inserted_at_)
    else:
      print("Verificar parametro")
      sys.exit()
 contador=0
 # leitura do conteudo extraido do kafka
 for message in consumer:
    linha=message.value
    dict str = linha.decode("utf-8")
    mydata = ast.literal_eval(dict_str)
  # Carregando os registros no postgres
    verifica layout arquivo()
    cur = conn.cursor()
    cur.execute(postgres_insert_query, record_to_insert)
    cur.close()
    conn.commit()
    contador = contador + 1
 print("Registros carregados: ",contador)
consumer.close()
conn.close()
```

5) Apache Spark – Para o processamento em escala, foi implementado um cluster Spark, que executa o programa python, assim o processamento será distribuído no cluster.

```
spark-submit --master yarn consumer_final.py cash_outs_created spark-submit --master yarn consumer_final.py cash_outs_processed spark-submit --master yarn consumer_final.py cash_outs_refunded spark-submit --master yarn consumer_final.py cash_ins
```



6) PowerBI – Criado um dashboard e um relatório no PowerBI, com opções de filtros diversos.



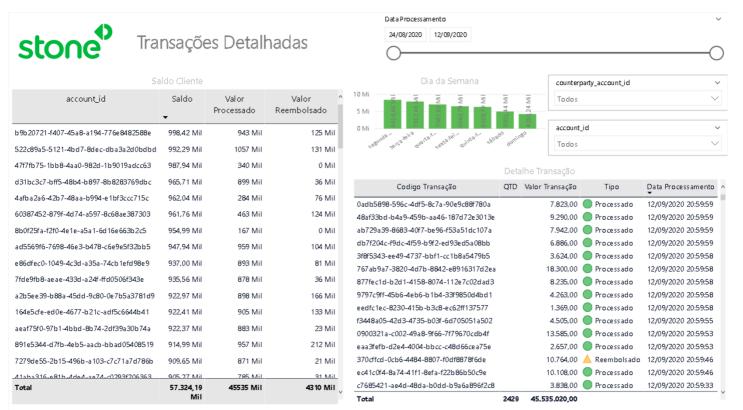


Figura 2 - Relatório detalhado, com diversas opções de filtros



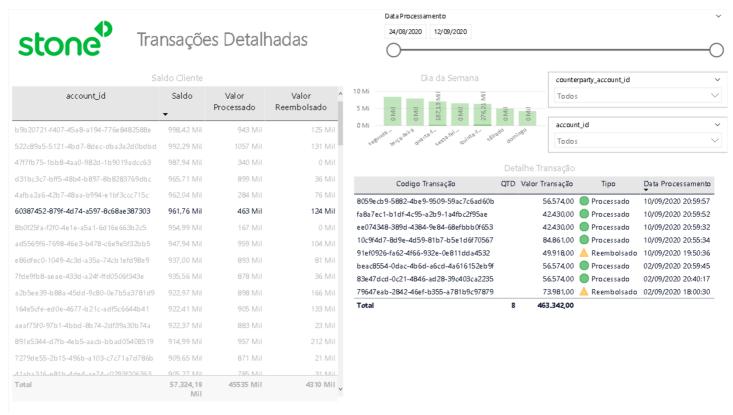


Figura 3 - Na tabela a esquerda, com informações do saldo, valor processado e valor reembolsado por cliente, basta escolher um cliente, todos os dados são filtrados para esse cliente. Do lado direito, é mostrado o detalhamento das transações do cliente selecionado.

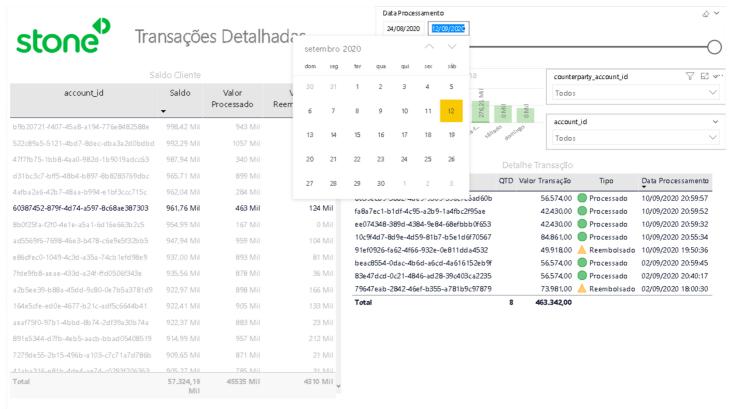


Figura 4 - Opção de filtro por data, pode ser uma data específica ou um intervalo de datas.



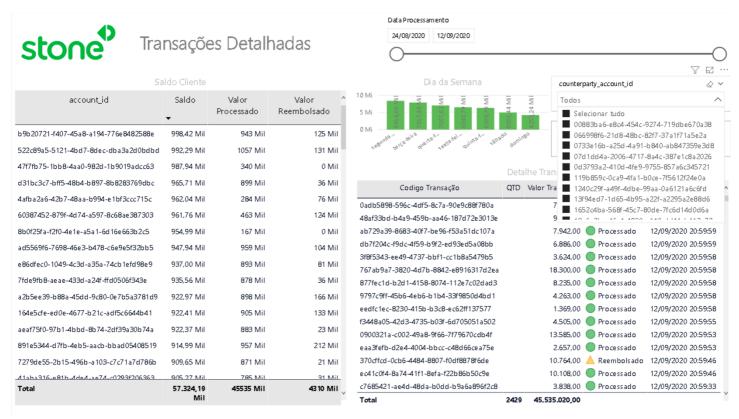


Figura 5 - Opção de filtro por counterparty.

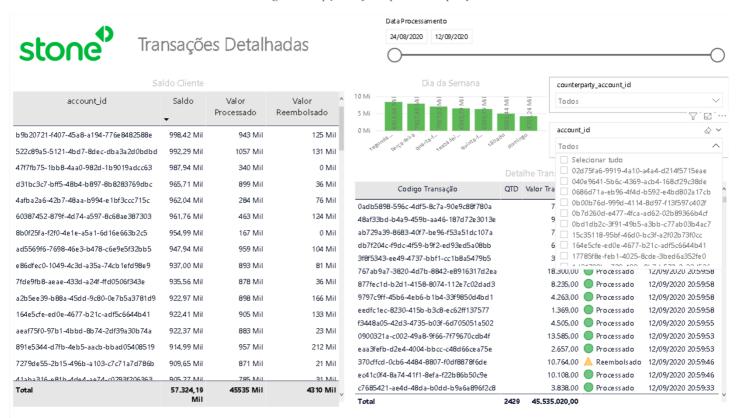


Figura 6 - Opção de filtro por account_id.



6 – Conclusão

Documento feito com o propósito de dar uma noção macro do funcionamento, de como é feito o processamento e como foi utilizado cada ferramenta.