

You **Tube**



**Spotify**

amazon **MP3**

**PANDORA**



COLLECTING COPYRIGHTS MASSIVELY WITH BIG DATA

# BIG DATA & SAS

- **O que é Big Data?**

- Big Data e Hadoop são sinónimos.
- É um conjunto de aplicações/serviços aplicativos distribuídos.

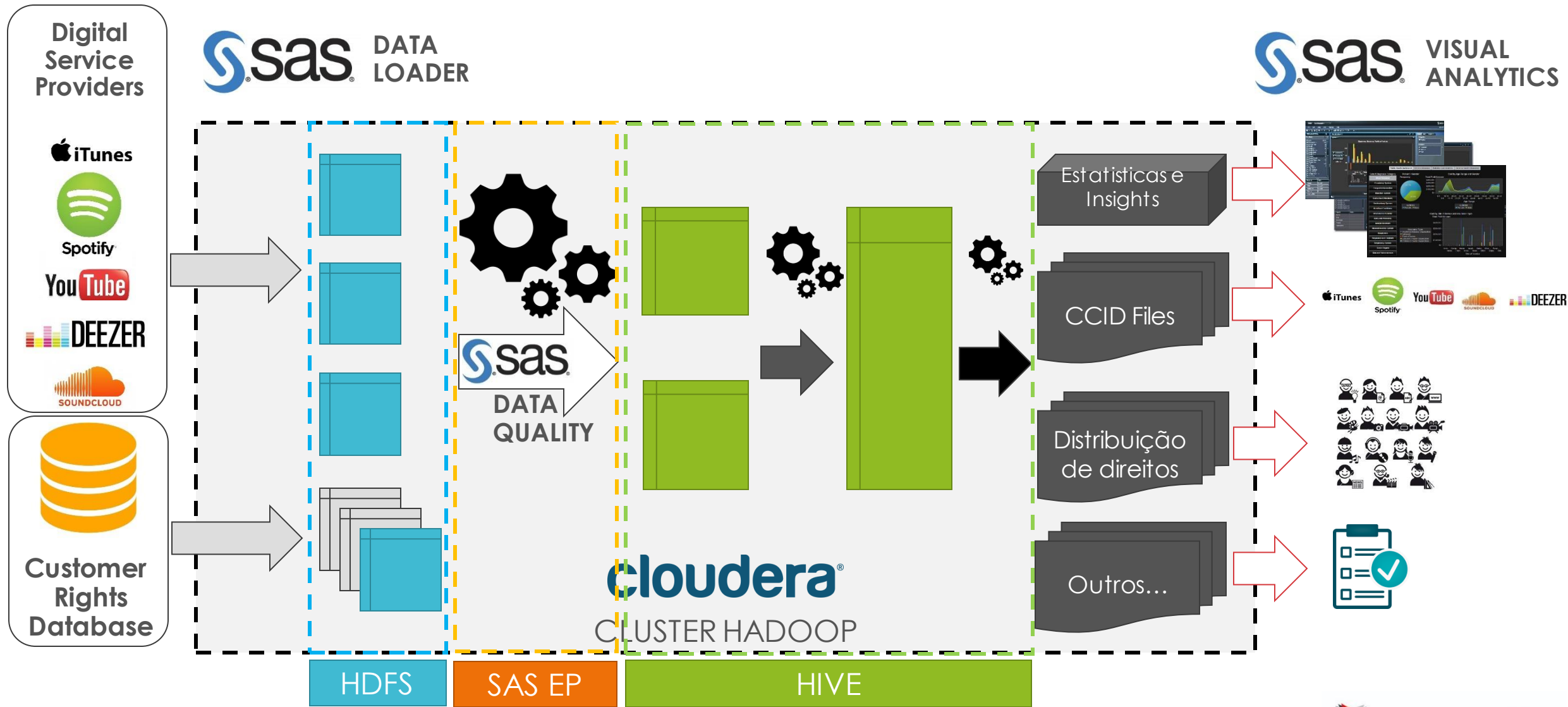
- **Para que serve e quais são as vantagens face ao que já existe?**

- Para processar volumes de informações, nos seus formatos nativos, uma vez que a sua principal vantagem é a capacidade de leitura e escrita (face a um sistema tradicional)

- **Como se usa de uma forma prática?**

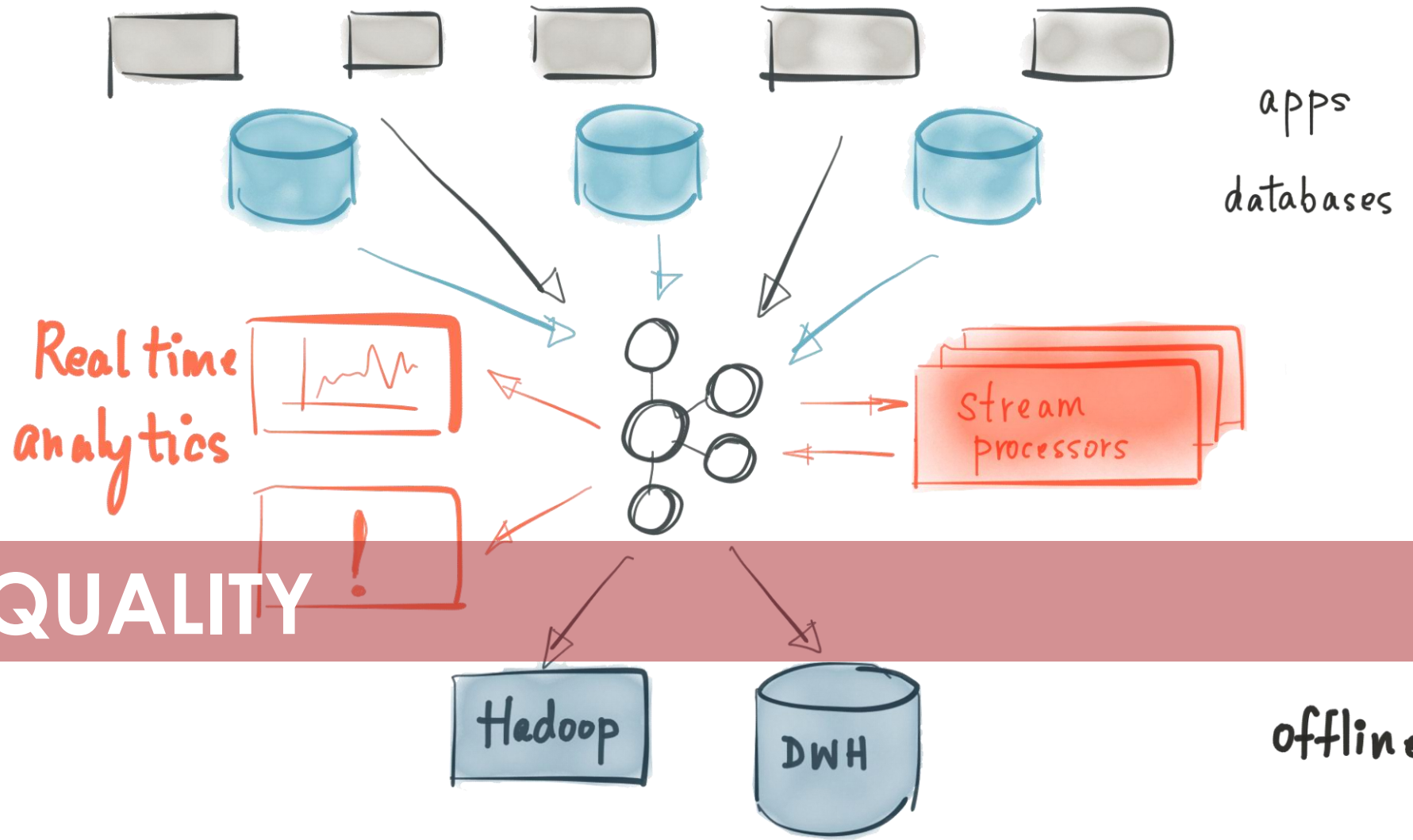
- Através de ferramentas visuais como o **SAS Data Loader** (para carregamento e processamento), **SAS Data Quality** (para melhorar a informação) e o **SAS Visual Analytics** (para reporting e análise).

# ARQUITETURA



COLLECTING COPYRIGHTS MASSIVELY WITH BIG DATA

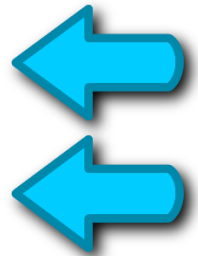
# STREAM DATA PLATFORM



**DATA QUALITY**

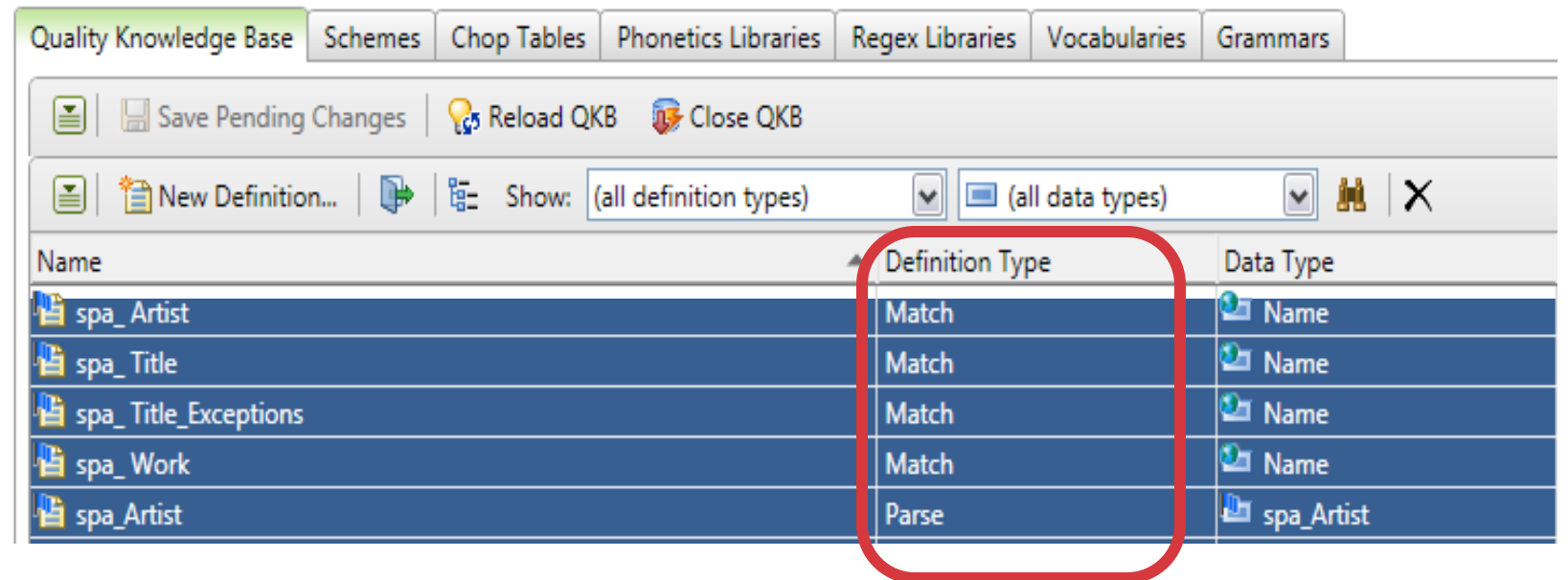
# DATA QUALITY - DESAFIO

- Matching de *Obras e Artistas* dos Digital Service Providers na base de dados de referência



# DATA QUALITY - ABORDAGEM

- Configuração SAS Data Quality Knowledge Base
- Definições de
  - *Parsers*
  - *Match Codes*
- Aplicado a
  - *Autores*
  - *Artistas*
  - *Obras*
  - *Títulos*



The screenshot shows the SAS Data Quality Knowledge Base interface. The 'Quality Knowledge Base' tab is active, and the 'Schemes' sub-tab is selected. The interface includes a toolbar with options like 'Save Pending Changes', 'Reload QKB', and 'Close QKB'. Below the toolbar, there are filters for 'Show: (all definition types)' and '(all data types)'. The main area displays a table with the following data:

Name	Definition Type	Data Type
spa_Artist	Match	Name
spa_Title	Match	Name
spa_Title_Exceptions	Match	Name
spa_Work	Match	Name
spa_Artist	Parse	spa_Artist



# DATA QUALITY - UTILIZAÇÃO

- Utilizado através de:
  - Reference Database
  - SAS Data Loader
- Os Match Codes obtidos são usados nos critérios de cruzamento de dados

SAS® Data Loader

PARSE DATA *rec\_artists: artistname1, artistname2, artistname3, artistname4, artistnameothers, artistname5*

GENERATE MATCH CODES  
*artistname1\_match\_code, artistname2\_match\_code, artistname3\_match\_code, artistname4\_match\_code, artistname5\_match\_code, artistnameothers\_match\_code, rec\_title\_codif\_match\_code, rec\_title\_codif\_mc\_exc*

Select the columns that you want to generate match codes for, the definition you want to apply, the sensitivity of the match and enter a .

[Return to Transformations](#)

Locale:  
Portuguese (Portugal) [Select a different locale](#)

Column:	Definition:	Sensitivity:	New Column Name:	
<a>▲</a> artistname1	spa_Artist	95 (high)	artistname1_match_code	✕
<a>▲</a> artistname2	spa_Artist	95 (high)	artistname2_match_code	✕
<a>▲</a> artistname3	spa_Artist	95 (high)	artistname3_match_code	✕
<a>▲</a> artistname4	spa_Artist	95 (high)	artistname4_match_code	✕
<a>▲</a> artistname5	spa_Artist	95 (high)	artistname5_match_code	✕
<a>▲</a> artistnameothers	spa_Artist	95 (high)	artistnameothers_match_code	✕
<a>▲</a> rec_title_codif	spa_Title	95 (high)	rec_title_codif_match_code	✕
<a>▲</a> rec_title_codif	spa_Title_Exceptions	95 (high)	rec_title_codif_mc_exc	✕

[+ Add Column](#)

[Next](#) [Add Another Transformation](#)

# DATA QUALITY - EXEMPLO

Reference Database



Rod Stewart	Tonight's the Night (Gonna Be Alright)	Rod Stewart Tom Dowd
4B~2\$&G6F^	9GDT\$\$\$VB\$\$\$\$	4B~2\$&G6F^ 7HB\$\$\$D^~J\$\$\$

Data Loader



Cleanse Data  
Cleanse data in Hadoop by performing data quality transforms

Rod Stewart	Tonight Night	Rod Stewart Tom Dowd
4B~2\$&G6F^	9GDT\$\$\$VB\$\$\$\$	4B~2\$&G6F^ 7HB\$\$\$D^~J\$\$\$

DSP Data



Rod Stewart	Tonight Night	Rod Stewart feat Tom Dowd
-------------	---------------	---------------------------



**Browse Tables**

Browse tables or open a table to see its contents

**Chain Directives**

Run multiple directives in a specific order

**Run Status**

Show the status of current and previous directive executions

**Saved Directives**

Open a previously created directive to run, view or edit

**Cluster-Survive Data**

Define rules to cluster similar records into groups and optionally create a best record to ...

**Delete Rows**

Delete rows from a selected table. Requires Hive 14 or above.

**Match-Merge Data**

Match-merge rows from one or more source tables into a single row and output a single ...

**Query or Join Data**

Query a table, or join data from multiple tables

**Run a Hadoop SQL Program**

Run custom SQL code

**Run a SAS Program**

Run custom SAS code

**Sort and De-Duplicate ...**

Query, sort, or de-duplicate the data in an existing Hadoop table

**Transform Data**

Transform data from a Hadoop table

# SAS DATA LOADER

**Transpose Data**

Transpose data from a Hadoop table

**Copy Data from Hadoop**

Copy Data from Hadoop into a database

**Copy Data to Hadoop**

Copy data from a database into Hadoop

**Import a File**

Import data from a file into Hadoop

**Load Data to LASR**

Copy data from a source and load it into LASR. Existing data in the target table will be re...

**Cleanse Data**

Cleanse data in Hadoop by performing data quality transforms

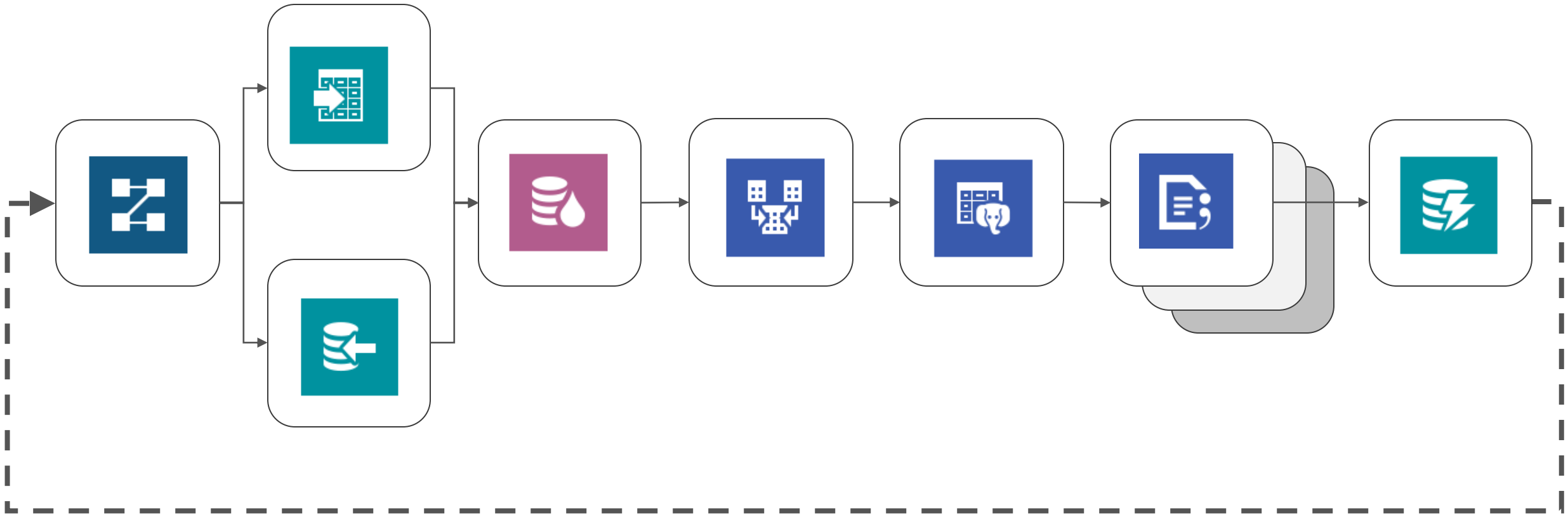
**Profile Data**

Generate a profile report of the data in a table

**Saved Profile Reports**

Explore previously generated profile reports

# SAS DATA LOADER



ES

graph



stats

PROCESSES



# ANALYTICS



model

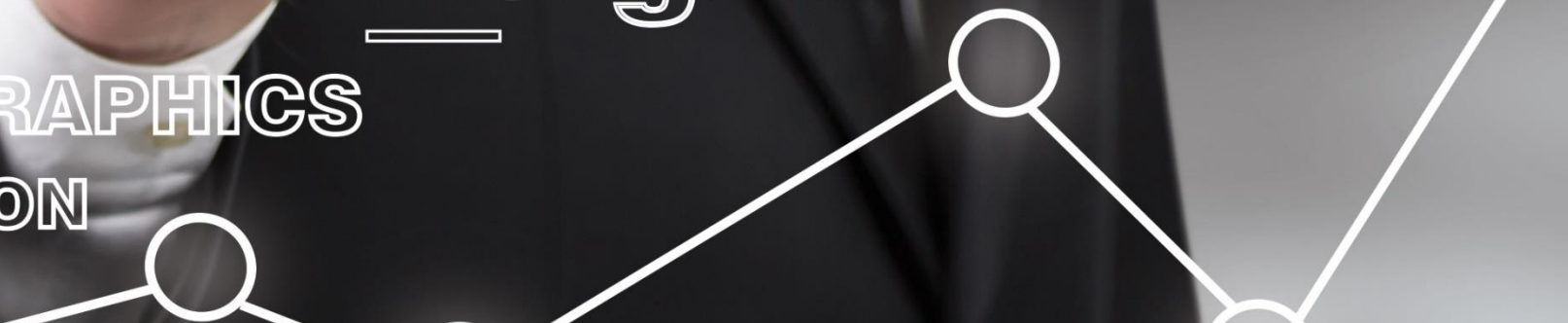
# insight

OPPORTUNITIES

GRAPHICS

PREDICTION

BUS



# VANTAGENS

- O Hadoop é uma plataforma de armazenamento altamente escalável, rápido e de baixo custo;
- SAS Data Loader permite que os utilizadores funcionais acessem facilmente ao Hadoop;
- SAS Data Quality potencia o *matching* dos dados provenientes dos DSPs na base de dados de referência;
- SAS Visual Analytics fornece *insights* e estatísticas sobre a informação;
- Os processos IT de negócio reduzem drasticamente o tempo de processamento;

# PORQUÊ A TIMESTAMP:BIW?



11 anos dedicados a projetos em Data Warehouse, BI & Analytics e EPM



Experiência em Big Data



Metodologias de desenvolvimento para soluções Big Data



Soluções Hadoop suportados em ambientes em Produção



Consultoria especializada em soluções Big Data

# POSICIONAMENTO FACE AO BIG DATA

## Transformando

A informação armazenada permite gerar análises preditivas e novos fluxos de valor para o negócio



## Otimizando

O desempenho operacional é otimizado com a utilização de Big Data, melhorando processos, tempos e resultados



## Explorando

O potencial da utilização de Big Data leva a experiências Hadoop e resultados básicos, mas sem efeitos práticos e reais no negócio



## Informado

Big Data é um termo conhecido, mas não se reflete em estratégias de negócios ou processos





# ASK BIGGER QUESTIONS

E VOCÊS, EM QUE ESTÁGIO DE ADOÇÃO SE  
ENCONTRAM?





OBRIGADO

