

ETL com Python: Luigi

Atividade 2

Equipe 8:

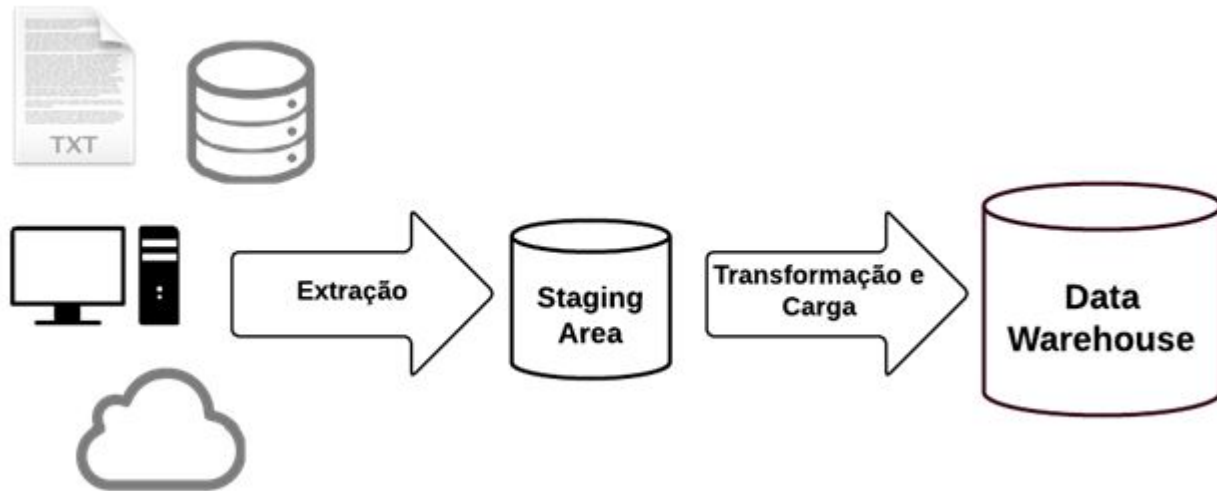
Gian Giovanni Rodrigues, Jessyca Moraes,
Tammy Gusmão, Thalita Alves

O que é o processo ETL?

- Muito popular quando o assunto é *data warehouse* e Business Intelligence (BI).
- Seu nome é um acrônimo para *Extract*, *Transform* e *Load*, que são as três etapas que constituem o método.
- Principal propósito do seu uso: combinar dados de diversas fontes para construir um *data warehouse*.
 - Melhora na performance do processamento do banco de dados.
 - Visualização compreensível dos dados.
 - Migração de dados de um sistema a outro.

O que é o processo ETL?

- Metodologia que consiste em três etapas bem definidas:



O que é o processo ETL?

- **Sobre a etapa de extração:**
 - Dados são retirados de um sistema-fonte.
 - Dados são convertidos para um formato único;
- **Sobre a etapa de transformação:**
 - Dados retirados passam por :
 - limpeza, correção e padronização,
 - tratamento de desvios e inconsistências;
 - Dados devem estar de acordo com as regras de negócio!
- **Sobre a etapa de carregamento:**
 - Dados são carregados no local apropriado (outro sistema);
 - Etapa define a característica de persistência dos dados.

Vantagens do uso do ETL

- Como vantagens, podemos citar:
 - Facilita a análise e geração de relatórios sobre dados relevantes, aumentando a produtividade dos profissionais analíticos;
 - Provê histórico completo para a empresa (via *data warehouse* corporativo);
 - Suporta fluxos velozes de dados (*streaming data*), permitindo agir imediatamente de acordo com certa oportunidade, com base no que está acontecendo em dado momento.

Possibilidades com o ETL

- Possibilita trabalhar em conjunto com outras ferramentas de integração de dados e com outros aspectos de gerenciamento de dados:
 - Todo tipo de Big Data;
 - Acesso a dados self-service;
 - Compreensão de metadados;
 - Auxilia na manutenção da qualidade dos dados;
 - Etc.

Por que usar o ETL no processo de DS?

- Muitas técnicas e práticas utilizadas na integração de dados são empregadas também pelos cientistas de dados:
 - SQL para acesso e transformação dos dados de uma base;
 - Mapeamento de dados, que ocorre na etapa de transformação, fornece instruções detalhadas sobre como obter dados necessários para processar, além de descrever qual campo de origem é mapeado para qual campo de destino;
 - Scripts executados em planos de fundo, movendo e transformando dados;
 - Uso do *master data management* (MDM) com o propósito de unir os dados, oriundos de diversas fontes, para criar uma visão única deles;
 - Etc.

Luigi, uma ferramenta ETL

- Criada e usada pela Spotify;
- Tudo nele é programado em Python;
- Melhor indicado para lidar com processos de lote de longa duração, portanto, consegue lidar com tarefas que vão além do escopo do ETL;
- Permite criar pipelines complexas de ETL;
- Acompanhamento de tarefas de ETL através de um *dashboard* em interface *web*;
- Inclui suporte a trabalhos de MapReduce em Hadoop, Hive e Pig.

TASK FAMILIES

Clear selection

data_management

PartsReport

14 PartsReport.CombineReports

44 PartsReport.DatedBOReport

16 PartsReport.DownloadReport

16 PartsReport.ExtractReport

2 PartsReport.GenerateFiles

6 PartsReport.LoadTable

2 PartsReport.LoadPartPlant

2 PartsReport.LoadPartPlantUse

2 PartsReport.LoadParts









2 PartsReport.DebateScrum

PFEP

BillOfMaterials

Others

1 LoadTable

**PENDING TASKS**
3**RUNNING TASKS**
1**BATCH RUNNING TASKS**
0**DONE TASKS**
134**FAILED TASKS**
3**UPSTREAM FAILURE**
1**DISABLED TASKS**
0**UPSTREAM DISABLED**
0

Displaying tasks of family LoadTable.

Show 10 entries

Filter table: Filter on Server ☐

	Name	Details	Priority	Time	Actions	
	RUNNING	LoadTable	table=FUP, date=2019-05-28	0	5/28/2019, 8:00:17 AM	

Showing 1 to 1 of 1 entries (filtered from 142 total entries)

Previous 1 Next

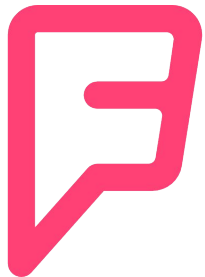
Luigi, uma ferramenta ETL

- As tarefas podem ser de qualquer natureza e pode incluir bibliotecas como Hadoop, Spark, Hive, etc.
- A ferramenta vem com uma *toolbox* que ajuda o usuário a criar tarefas, além de ter *templates* (exemplos) de tarefas mais comuns.
- Possibilita o acesso a HDFS (sistema de arquivos escalonáveis e distribuído).
 - Todas as operações com estes arquivos são atômicos, portanto não ocorre *crash* da *pipeline* de dados.

Luigi: vantagens x desvantagens

Vantagens	Desvantagens
Facilita gerenciamento de processos de lote.	Não lida com processos contínuos de <i>streaming</i> .
Permite incluir tarefas de diferentes bibliotecas/pacotes de software (Hadoop, Hive, Pig, Cascading).	Baixa escalabilidade se comparado ao Airflow.
Grafos de dependência do <i>workflow</i> (DAGs) são especificados dentro do Python, facilitando a construção de um grafo complexo de dependências.	O <i>workflow</i> pode afetar fatores externos ao Python. Por exemplo: scripts em execução do Pig, arquivos externos sendo acessados.
Não há configuração em arquivos XML ou arquivos externos.	Precisa programar o código de acesso a base de dados (não há <i>templates</i>).

Empresas que adotaram o Luigi



Red Hat



linx
+ neemu
+ chaordic

GetNinjas 


Hotels.com

Exemplo de Processo ETL

Feito no Google Colaboratory
com a biblioteca Pandas

Consultas realizadas

1. Ranking de cidades, de destino da compra, de acordo com o total de vendas.
2. Ranking de países de acordo com a quantidade e total de vendas realizadas.
3. Classificação de usuário por quantidade de compra.

1

	ship_city	total_de_vendas
0	Cunewalde	110277.305030
1	Graz	104874.978144
2	Boise	104361.949540
3	Rio de Janeiro	51956.979901
4	Albuquerque	51097.800828
5	Cork	49979.905081
6	Sao Paulo	40486.461552
7	London	39434.359953
8	Brandenburg	30908.383873
9	Bräcke	29567.562490
10	Montréal	28872.190156
11	Seattle	27363.604900
12	München	26656.559404
13	Luleå	24927.577431
14	Charleroi	24088.780281

2

	country	nro_vendas	u\$_total
0	USA	352	245585.0
1	Germany	328	230285.0
2	Austria	125	128004.0
3	Brazil	203	106926.0
4	France	184	81358.0
5	UK	135	58971.0
6	Venezuela	118	56811.0
7	Sweden	97	54495.0
8	Canada	75	50196.0
9	Ireland	55	49980.0
10	Belgium	56	33825.0
11	Denmark	46	32661.0
12	Switzerland	52	31693.0
13	Mexico	72	23582.0
14	Finland	54	18810.0
15	Spain	54	17983.0
16	Italy	53	15770.0
17	Portugal	30	11472.0
18	Argentina	34	8119.0
19	Norway	16	5735.0
20	Poland	16	3532.0

3

	level	customer_id		company_name	contact_name	contact_title	address	city	region	postal_code	country	phone	fax
0	Padrão	ALFKI		Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	None	12209	Germany	030-0074321	030-0076545
1	Basic	ANATR	Ana Trujillo Emparedados y helados		Ana Trujillo	Owner	Avda. de la Constitución 2222	México D.F.	None	05021	Mexico	(5) 555-4729	(5) 555-3745
2	Padrão	ANTON		Antonio Moreno Taquería	Antonio Moreno	Owner	Mataderos 2312	México D.F.	None	05023	Mexico	(5) 555-3932	None
3	Padrão	AROUT		Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	None	WA1 1DP	UK	(171) 555-7788	(171) 555-6750
4	Super	BERGS		Berglunds snabbköp	Christina Berglund	Order Administrator	Berguvsvägen 8	Luleå	None	S-958 22	Sweden	0921-12 34 65	0921-12 34 67
...
84	Super	WARTH		Wartian Herkku	Pirkko Koskitalo	Accounting Manager	Torikatu 38	Oulu	None	90110	Finland	981-443655	981-443655
85	Padrão	WELLI		Wellington Importadora	Paula Parente	Sales Manager	Rua do Mercado, 12	Resende	SP	08737-363	Brazil	(14) 555-8122	None
86	Padrão	WHITC		White Clover Markets	Karl Jablonski	Owner	305 - 14th Ave. S. Suite 3B	Seattle	WA	98128	USA	(206) 555-4112	(206) 555-4115
87	Padrão	WILMK		Wilman Kala	Matti Karttunen	Owner/Marketing Assistant	Keskuskatu 45	Helsinki	None	21240	Finland	90-224 8858	90-224 8858
88	Padrão	WOLZA		Wolski Zajazd	Zbyszek Piastrzyeniewicz	Owner	ul. Filtrowa 68	Warszawa	None	01-012	Poland	(26) 642-7012	(26) 642-7012

89 rows × 12 columns

Mais detalhes no .ipynb