Extraindo e transformando dados de uma API pública

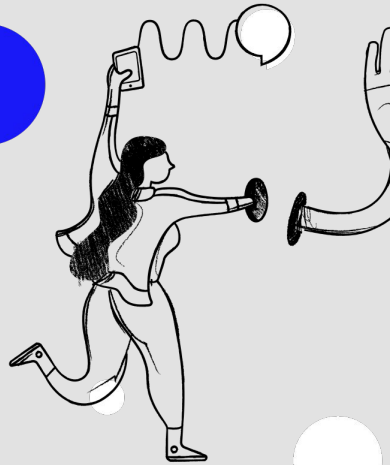
Marcos Camargo

02 de Outubro de 2020

WAVY

Marcos Camargo

- Engenheiro de dados na Wavy
- Cientista/Engenheiro de dados na Birdie
- BCC 015
- Semcomp 19, 20 e 21



Agenda

- Apresentação
- A área de dados
- Engenharia de dados
- Material de apoio
- Exemplo prático

Wavy Global

WAVY

+1.5BI

MENSAGENS ENVIADAS POR MÊS

+200K

USUÁRIOS ÚNICOS MENSAIS

+70

OPERADORAS CONECTADAS

200 Terabytes

Data lake

3 Terabytes

Crescimento Mensal

~50 Servidores

Bancos de dados de produção

O que são dados?

Tweet

https://twitter.com/semcomp/status/1308920925631455233


```
{"created_at": "Thu Sep 24 00:07:21 +0000 2020",
  "id": 1308920925631455233,
  "id_str": "1308920925631455233",
  "text": "Saiu o cronograma da Semcomp 23!\n\nPara se inscrever, basta preencher o formulário em https://t.co/OWx6gjQR4z https://t.co/UWpO42TZgf",
  "truncated": false,
  "entities": {"hashtags": [],…
      "small": {"w": 680, "h": 680, "resize": "fit"}}}]},
  "extended_entities": {"media": [{"id": 1308920917481971713,…
      "small": {"w": 680, "h": 680, "resize": "fit"}}},
    {"id": 1308920917481918465,…
      "small": {"w": 680, "h": 680, "resize": "fit"}}},
    {"id": 1308920917486194689,…
      "small": {"w": 680, "h": 680, "resize": "fit"}}},
    {"id": 1308920917477777409,…
      "large": {"w": 1000, "h": 1000, "resize": "fit"}}}]},
  "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {"id": 16714079,
    "id_str": "16714079",
    "name": "Semcomp 23",
    "screen_name": "semcomp",
    "location": "ICMC - USP, São Carlos",
    "description": "Twitter oficial da Semana da Computação - evento de tecnologia que ocorre anualmente no ICMC/USP em São Carlos, SP. Já nos preparativos da 23ª edição!",
    "url": null,
    "entities": {"description": {"urls": []}},
    "protected": false,
    "followers_count": 425,
    "friends_count": 275,
    "listed_count": 9,
    "created_at": "Mon Oct 13 00:34:43 +0000 2008",
```

Alguns tipos de dados

- Tabela

created_at	id	username	text	retweet_count	favorite_count
2020-09-24 00:07:21	13089209256314552	semcomp	Saiu o cronograma ...	20	47

- Comma-Separated Values - CSV

```
created_at,id,username,text,retweet_count,favorite_count
2020-09-24 00:07:21 +0000 ,1308920925631455233,semcomp,"Saiu o cronograma ...",20,47
```

- LOGs

```
[2020-09-24 00:07:21 +0000] - username semcomp tweeted "Saiu o cronograma ..." tweet_id 130892...
[2020-09-24 00:07:21 +0000] - tweet_id 1308920925631455233 retweet_count 20
[2020-09-24 00:07:21 +0000] - tweet_id 1308920925631455233 favorite_count 47
```

- Existem outros formatos mais específicos, binários, etc...

Onde esses dados podem estar?

Logs

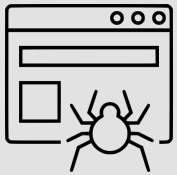
Banco de dados

elasticsearch

kafka

Mensageria / Eventos

Crawlers



PostgreSQL

APIs

Como extrair valor dos dados?

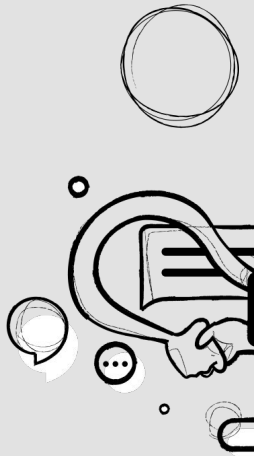
- Ciência de dados
  - Análise de sentimentos

- Análise de dados
  - Medir indicadores chaves
    - Novos usuários mensais
  - Responder perguntas baseada em dados
    - Qual a popularidade de cada perfil?
  - Insights
    - 70% dos usuários utilizam o máximo de caracteres do tweet -> aumentar limite de caracteres

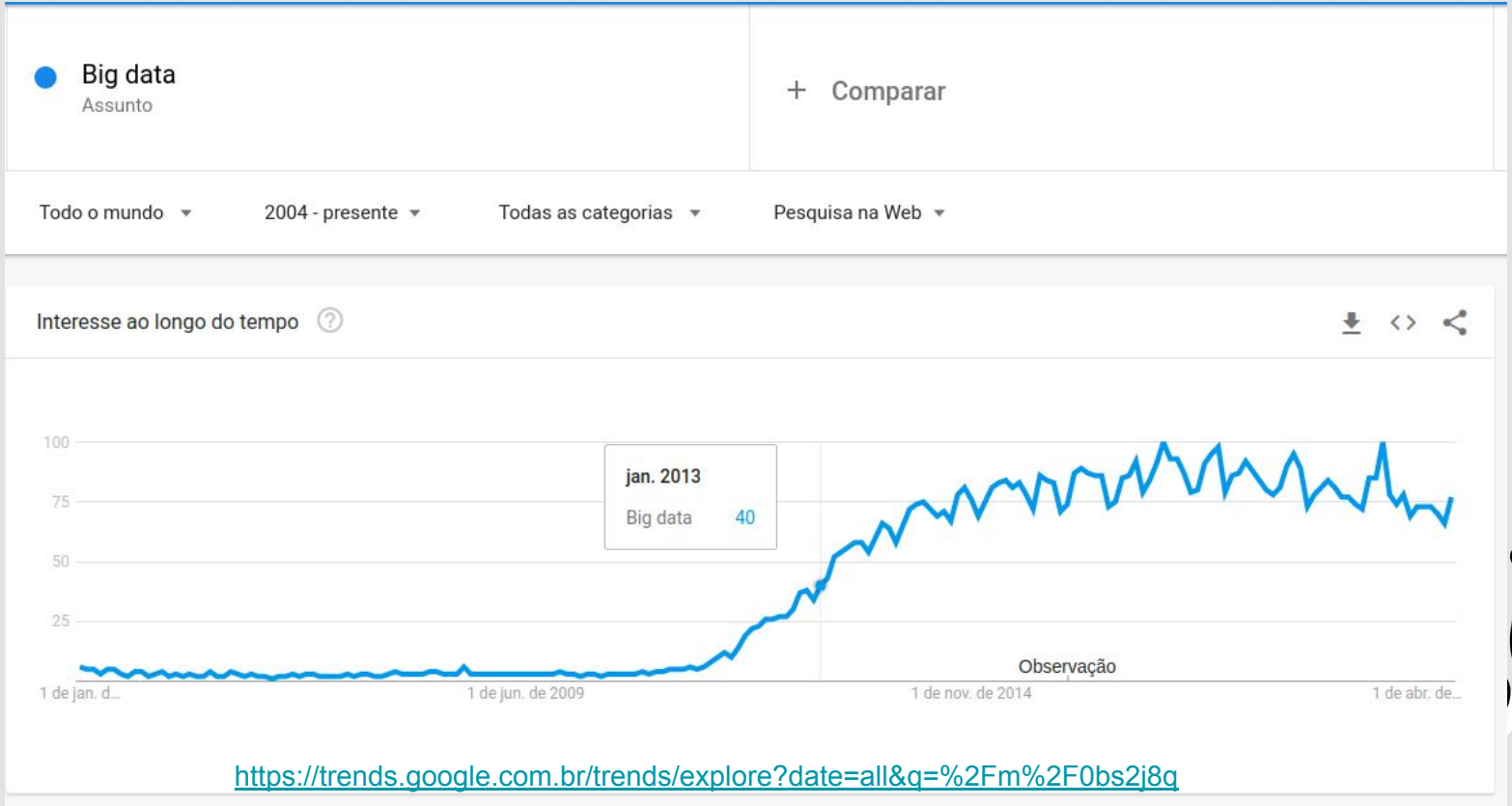
Um pouco sobre a área de dados

Organização de um time de Data & Analytics

- Não existe receita de bolo
- Grande leque de carreiras e possibilidades
- Diversos perfis de pessoas de dados dentro de uma empresa
- Muitas ferramentas, tecnologias, abordagens diferentes
- Explosão de dados: Big Data, Clusters



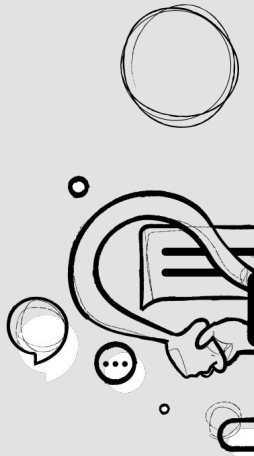
Google trends - assunto Big Data



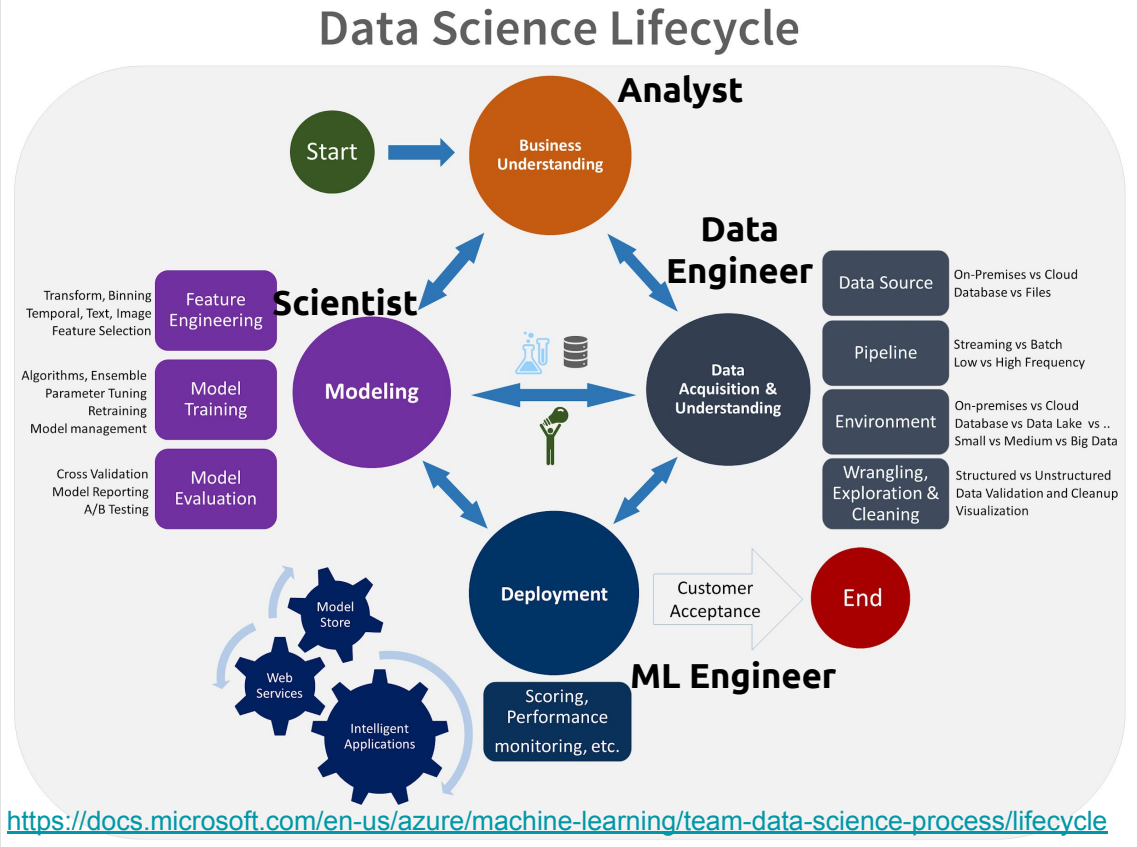
https://trends.google.com.br/trends/explore?date=all&q=%2Fm%2F0bs2j8q

Principais carreiras em dados

- Analista de dados
  - Responsáveis por analisar os dados, extraírem os padrões e responderem as perguntas do negócio
- Cientista de dados
  - especialistas em modelagem preditiva e aprendizado de máquina
- Engenheiro(a) de Machine Learning
  - responsável pela implementação e performance dos modelos
- Engenheiro(a) de dados
  - responsável pela aquisição, processamento e qualidade dos dados



Principais carreiras em dados



https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

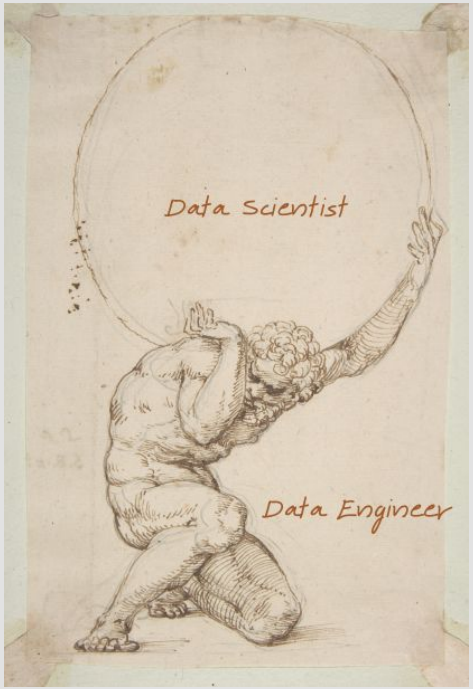
Introdução à engenharia de dados

O que é um Engenheiro(a) de dados ?

"Engenheiro(a) de Dados é um tipo especializado de Engenheiro(a) de Software que possibilita outros a responderem questões sobre grandes datasets com restrições específicas de latência e tempo." - **Nathan Marz**, Inventor do *Apache Storm* e da *Lambda Architecture*

Algumas responsabilidades

- Aquisição de dados
- Integração de diversas fontes de dados
- Criação de pipelines em batch e streaming
- Garantir a qualidade dos dados
- Garantir a segurança dos dados (LGPD)
- Arquitetura de sistemas distribuídos
- DevOps, DBA e Infra



ETLs

Extract, Transform and Load

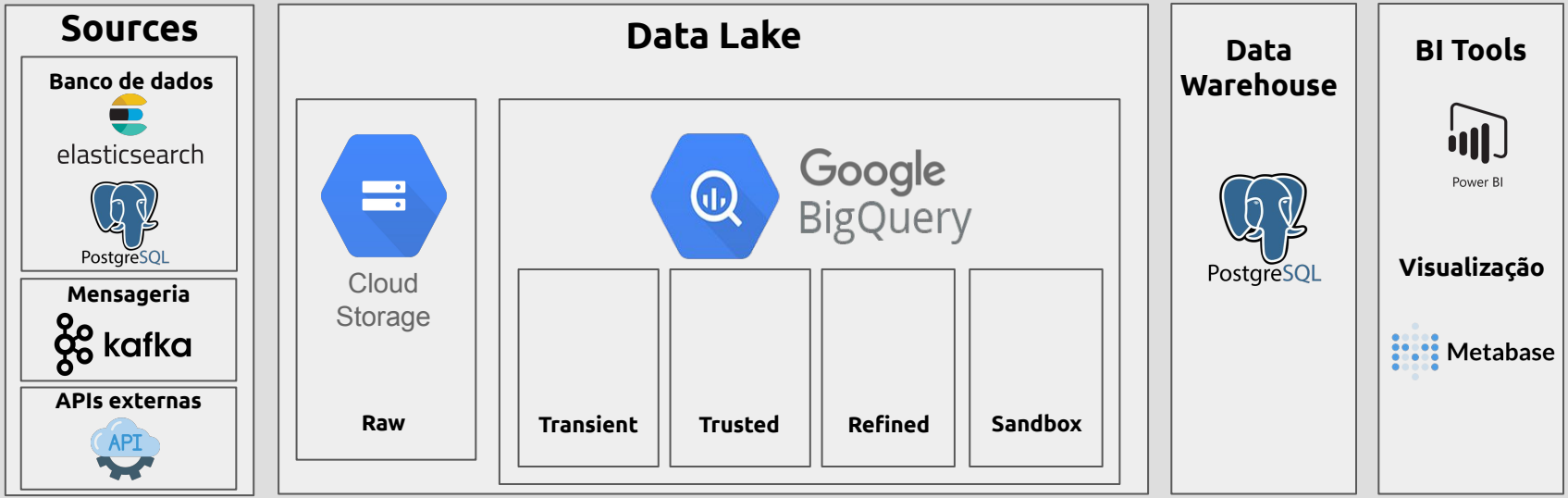
- **Extract**
  - Coleta de eventos
  - Extração de diversas bases de dados distintas
  - Ingestão via logs
- **Transform**
  - Combinação dos dados
  - Transformações (tipo, formato, limpeza) e estruturação
- **Load**
  - Armazenamento do dado no seu destino final
- Existe uma variação que é o **ELT**

Algumas ferramentas de ETLs

- **Processamento distribuído**

- **Orquestração**

Como é a estrutura de dados na Wavy



Top 5 buzzwords para por no CV

- **Data Lake**
  - Repositório que armazena conjuntos grandes e variados de dados brutos em formato nativo e de forma organizada
- **Data Warehouse**
  - Repositório de dados tratados e padronizados, geralmente utilizado para Business Intelligence
- **Data Visualization**
  - Representação visual dos dados, provêem uma forma acessível para entender tendências, outliers e padrões
- **Big Data**
  - Análise e interpretação de grandes volumes de dados, envolve processamento distribuído
- **Event Streaming**
  - Fluxo e processamento contínuo de dados, aplicações NRT, arquitetura de eventos

Dicas de projetos/aprendizado

Database Engineering | Data Pipeline

na iFood (Ver todas as vagas de emprego)

Campinas/Osasco/Remoto

As a **Database Engineer**, your challenges will be:

- Design, code, test, operate and solve production problems on services running on the cloud (specially AWS);
- Participate in product evolution prioritization, always looking to the best value gains to the business, basing your decisions on data;
- Face and solve scalability, maintainability and reliability challenges;

The ingredients we're searching for:

- Have strong knowledge of SQL, data architecture and relational data models and databases, performance tuning of PostgreSQL and queries;
- Have strong coding skills and architecture such as Java, Go, Ruby, Python;
- Have experience with infrastructure automation and configuration management (Chef, Terraform) with AWS EBS, ELB, ASG, MSK, EC2, RDS;
- Have strong experience with NoSQL Cassandra, ScyllaDB and DynamoDB;
- Have strong experience with Kafka;

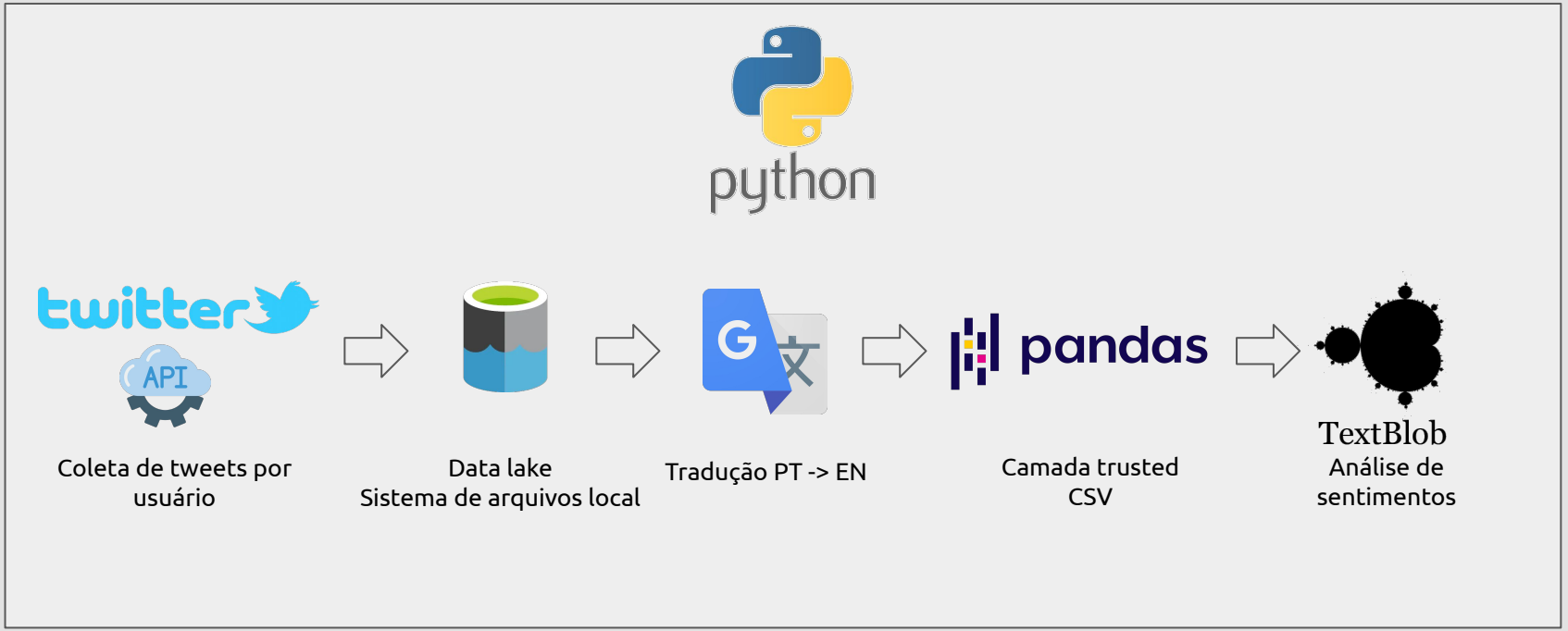
https://boards.greenhouse.io/ifood/jobs/4777039002

Material de apoio

- O que faz um engenheiro de dados? - Data Hackers
- Roadmap de estudos - Data hackers
- GitLab Data Engineering Handbook
- Comunidade Data Hackers (slack, newsletter, podcasts)
- Mediums: Towards data science - tag data engineering ...
- Postcasts Pizza de dados
- Databricks Community
- Public APIs: Twitter, Pokemon, AlphaVantage ...
- Public Datasets: Wikipedia, BigQuery open datasets ...

Vamos ver esses conceitos em prática?

Nosso case de estudo



Marcos Camargo
marcos.camargo@wavy.global
Linkedin: marcoscrcamargo