## 1 Etapas de um projeto de Big Data

Para que um projeto de Big Data tenha sucesso, algumas etapas devem ser seguidas. A depender do tipo de dado, essas etapas podem ser mais complexas.

### Coleta de Dados

Algumas empresas querem coletar todos os dados e depois verificar de que forma podem ser utilizados, mas o ideal é que os projetos de Big Data são direcionados por um objetivo bem específico, mesmo que depois vá se ampliando.

No mundo do comércio eletrônico, a quantidade de dados a serem coletados é bem vasta, cliques de *mouse,* localização, tipo de dispositivo, histórico de compras, são exemplos. No caso dos Ministérios Públicos, a maior variedade vem de dados não estruturados advindos de apreensões.

Para o compartilhamento de dados com os MPs, alguns órgãos usam FTP, Webservice, Dados abertos, ou mesmo mídia física. Independentemente do canal de comunicação, uma preocupação importante é com a atualização desses dados. Sempre que possível, automatize a coleta dos dados. Para automatização, pode ser necessário utilizar alguma ferramenta de ETL. Pode-se usar o Pentaho Data Integrator, ferramenta gratuita.

A coleta de dados é a base para todo o trabalho.

### Limpeza dos dados

Esta etapa é considerada por muitos a mais importante do processo, e a mais trabalhosa. Nesta etapa, são identificadas anomalias ou discrepâncias que podem comprometer as análises. Dados anômalos (valores nulos, inconsistentes, mascarados, duplicados, etc.) serão removidos ou tratados. Também pode ser feito o enriquecimento dos dados, que é agregar dados de outras fontes ao conjunto de dados. Um problema muito comum com bases de dados são os campos data, por exemplo, data do fato de um processo sempre é algo passado, se uma base de processos tiver algum dado futuro,

esse dado apresenta erro.

Muitas bases de dados fornecidas por órgãos públicos usam como chave o CPF ou CNPJ. Por isso, é importante que os dados sejam normalizados (deixar com mesmo tamanho e formato, por exemplo) antes de adicionar a base ao Data Lake. No caso do CPF, pode ter bases que não colocam os zeros à esquerda, ou mesmo que utilizem o dado formatado. É recomendável adotar um padrão. A sugestão é que, para o CPF, sejam utilizados 11 dígitos, com zeros à esquerda, e sem formatação. E para o CNPJ, 14 dígitos.

Quando estamos tratando de dados não estruturados, vindos dos HDs apreendidos, devemos retirar arquivos duplicados ou desnecessários, como arquivos executáveis, bibliotecas de sistemas, *drivers* de dispositivos, por exemplo. Como selecionar os arquivos inúteis?

Existem listas de arquivos conhecidos, Known File Filter, que é um banco de dados com arquivos catalogados com seus respectivos *hashs*. Assim, arquivos desnecessários são facilmente reconhecidos e podem ser desprezados. *Link* para *download* https://www.nist.gov/itl/ssd/software-quality-group/nsrl-download

Para ampliar o conhecimento acerca do tema, sugiro a leitura do *post* disponível em: https://escoladedados.org/2016/09/guia-quartz-para-limpeza-de-dados/

**Mineração de dados**

Pode ser definida como um processamento de dados para a identificação de padrões. Para isso, podem ser usados diversos métodos, como estatística, expressões regulares, ou mesmo inteligência artificial.

Nessa fase, é possível, por exemplo, identificar que beneficiários de programas sociais estão fazendo doações eleitorais acima da capacidade financeira, sendo uma potencial irregularidade.

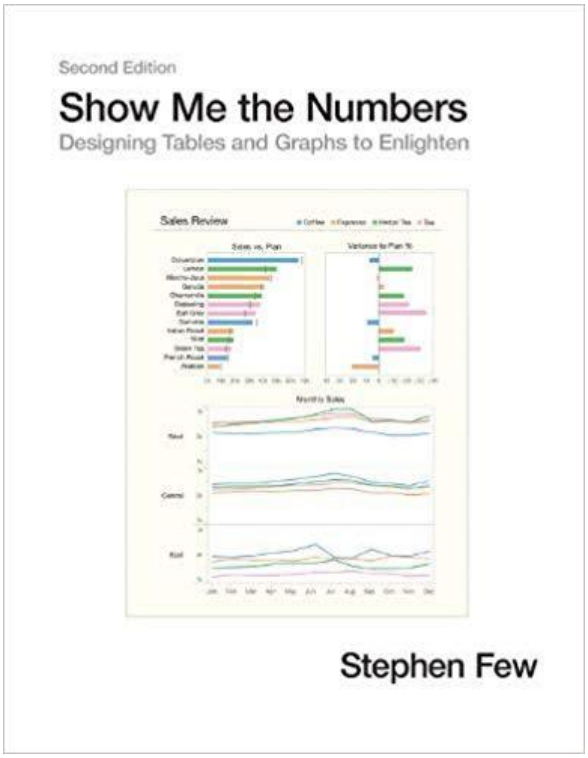
Para conhecer as técnicas de mineração de dados, leia o texto disponível em

https://www.ibm.com/developerworks/br/library/tecnicas-mineracao-de-dados/index.html

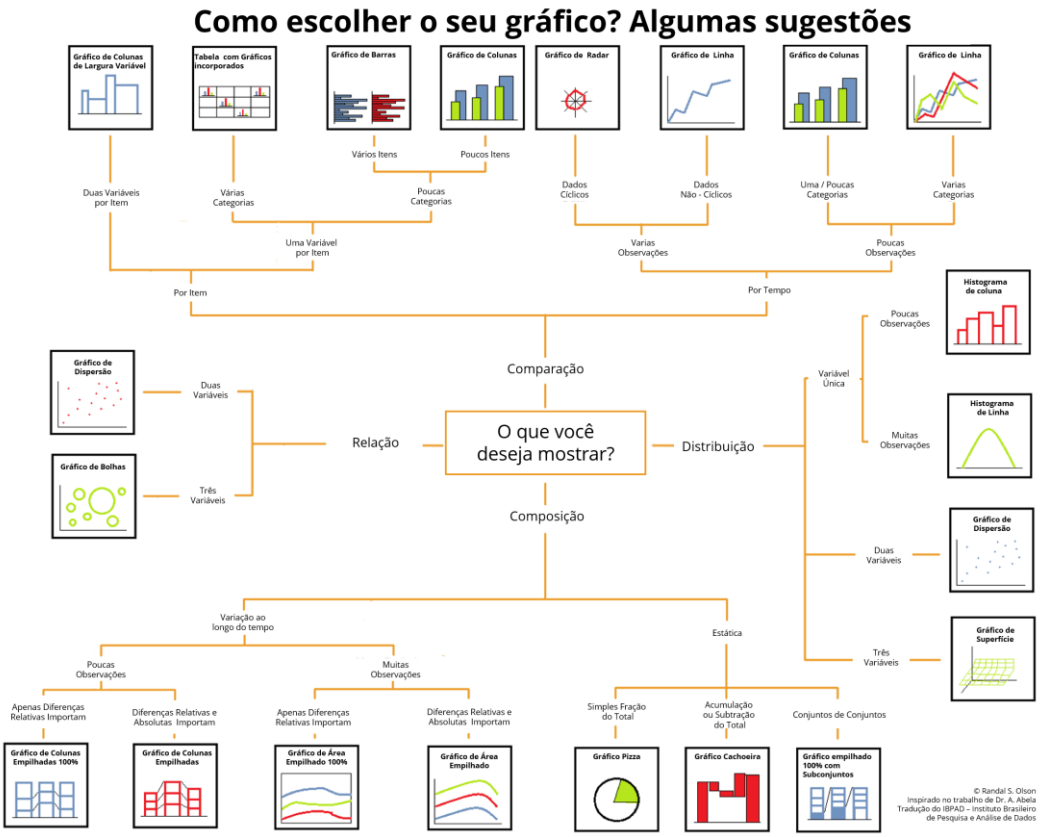
**Visualização de Dados**

Não basta ver uma grande quantidade de dados organizada em planilhas. O ideal é usar um meio de fácil interpretação para que as pessoas tenham acesso às informações encontradas. O uso de gráficos ou infográficos são ferramentas visuais de extrema importância para facilitar o entendimento. Gráfico não dever ser usado apenas para substituir uma planilha. Deve-se estudar a melhor forma de apresentar determinada informação.

O assunto é muito amplo. Por isso, recomendo a leitura do livro Show Me The Number, do renomado autor Stephen Few.




Existem diversos estudos que apresentam as melhores formas de apresentar os dados. Abaixo consta a apresentação traduzida pelo IBPAD do autor Randal S. Olson.




Fonte: https://www.ibpad.com.br/blog/analise-de-dados/qual-e-o-grafico-mais-adequado/

Conhecer o público-alvo é importante para avaliar os dados que buscarão interpretar. Muitas vezes, é necessário produzir várias visualizações de acordo com os vários públicos-alvo envolvidos.

Recomendo a leitura do *post* disponível em: https://clusterdesign.com.br/visualizacao-de-dados-ciencia-arte-ou-ambos/

## 2 Indexação de dados

Um dos principais usos das soluções de Big Data é para armazenar, processar e indexar dados estruturados e não estruturados. Para se indexar dados não estruturados, como documentos pdf, doc, xls dentre outros, é necessário, primeiramente, extrair o texto desses arquivos, também conhecido como extração de conteúdo. Para isso, existem diversas formas, de acordo com o tipo de arquivo.

Para facilitar esse trabalho, existe uma biblioteca chamada Apache Tika, que extrai o texto de vários formatos de arquivos, disponibilizado por meio de lib jar ou aplicação server via api. Também temos, no sistema linux, um aplicativo chamado pdftotext. Abaixo temos o exemplo de um pdf e sua extração.

DIÁRIO OFICIAL DA UNIÃO

Publicado em: 14/06/2019 | Edição: 114 | Seção: 3 | Página: 1

Órgão: Presidência da República/Casa Civil/Instituto Nacional de Tecnologia da Informação

AVISO DE REVOGAÇÃO

PREGÃO ELETRÔNICO Nº 2/2019

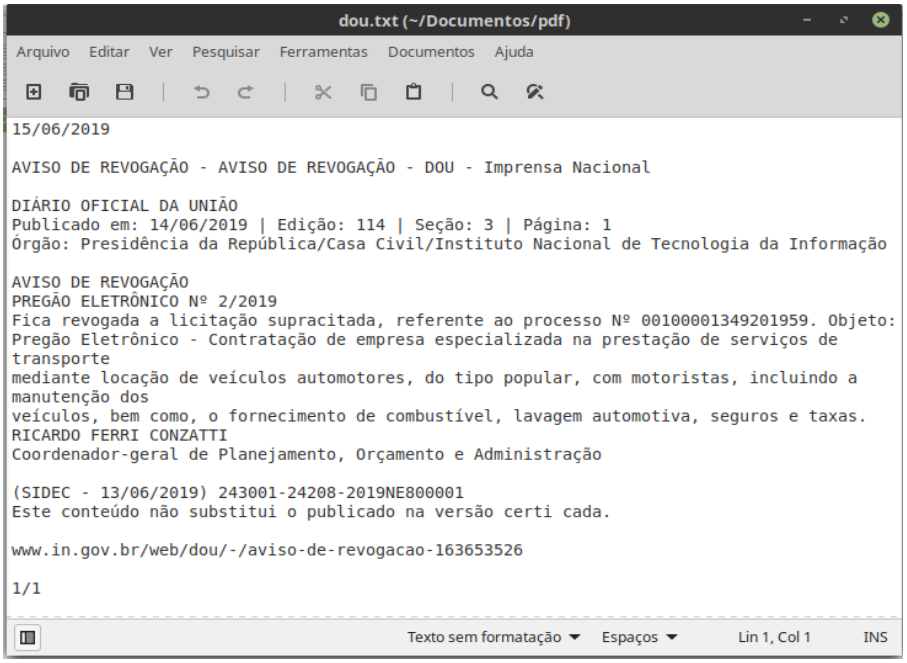
Fica revogada a licitação supracitada, referente ao processo Nº 00100001349201959. Objeto: Pregão Eletrônico - Contratação de empresa especializada na prestação de serviços de transporte mediante locação de veículos automotores, do tipo popular, com motoristas, incluindo a manutenção dos veículos, bem como, o fornecimento de combustível, lavagem automotiva, seguros e taxas.

RICARDO FERRI CONZATTI
Coordenador-geral de Planejamento, Orçamento e Administração

(SIDEC - 13/06/2019) 243001-24208-2019NE800001

Este conteúdo não substitui o publicado na versão certificada.

Utilizei a extração com o comando ***pdftotext <arquivo entrada> <arquivo de saida>***, segue o resultado:




Para extração com o apache tika, baixei a versão server e para rodar o serviço

basta executar ***java -jar tika-server-1.21.jar***. Feito isso, podemos solicitar a extração do conteúdo com o ***curl -T dou.pdf http://localhost:9998/tika >doutika.txt***, o resultado é similar.

Com a extração de conteúdo feita, pode-se indexar com alguma ferramenta. Eventualmente, podemos precisar extrair algumas informações desse conteúdo por meio de expressão regular ou algoritmos de Processamento de Linguagem Natural antes de indexar. Por exemplo, extrair endereços de *e-mail*, CPFs, CNPJs, telefones, nomes de pessoas, processo conhecido como extração de entidades. Isso é feito para criar campos específicos ao se fazer a indexação.

Principais recursos de buscas:

- **Busca textual:** possui diversos recursos de linguística para ampliar as possibilidades de pesquisa, como remoção de acentos e *stopwords*, ignorar letras maiúsculas ou minúsculas, sinônimos, dentre outros.
- **Busca geoespacial:** busca por coordenadas geográficas para pesquisar, por exemplo, os restaurantes da região.
- **Destaque de termos**: destaque do termo pesquisado no texto retornado.
- **Busca facetada**: busca guiada com agregação de dados em categorias.

Os principais indexadores são Apache Solr ou ElasticSearch, ambos utilizam o Apache Lucene, e também são conhecidos como ferramentas de Recuperação de Informação.

**ElasticSearch**

Para utiliza-lo, basta realizar o *download* https://www.elastic.co/, descompactar e executar o comando **bin/elasticsearch** (linux) ou **bin\elasticsearch.bat** (windows) e já estará pronto para ser executado. Não há interface. Toda iteração é por meio de API Rest e documentos JSON. O Elastic é *schemaless*, ou seja, não trabalha com *schema* definido, permitindo adicionar novos campos ou índices apenas por meio da API. O

ecossistema da Elastic, mais conhecido como Stack ELK, fornece uma ferramenta visual chamada kibana, que permite monitorar o Elastic, realizar consultas e construir visualizações interativas. Montar um *cluster* Elastic é bem simples. Basta subir duas instâncias em computadores diferentes que o serviço de *discovery* dele encontra os nós e monta o *cluster*. Também é possível configurar manualmente. Na parte prática vamos executar o Elasticserach e realizar alguns testes.

Para criar um índice no Elastic, bastar fazer uso da API, realizando um ***PUT /customer*** que criará o índice *customer*. Para indexar documentos, basta realizar o **PUT /customer/_doc/1 { "name": "John Doe" }** que indexará um documento com o campo *name* e valor Jon Doe.

Segue a equivalência dos bancos relacionais com o ElasticSearch:

| **Banco Relacional** | **ElasticSearch** |
|---|---|
| Banco de Dados/Schema | Índice |
| Tabela | Tipo (não utilizado na versão 7) |
| Registro | Documento |
| Campo | Campo |

Como o ElasticSearch não possui schema definido, é possível que cada documento tenha uma estrutura de dados diferente no mesmo índice. A limitação é que um campo só pode ter um único tipo de dados.

Recomendo assistir o vídeo de Introdução ao ElasticSearch disponível em https://www.elastic.co/pt/webinars/getting-started-elasticsearch?baymax=rtp&elektra=home&storm=sub1&iesrc=ctr e a Introdução ao Kibana, disponível em https://www.elastic.co/pt/webinars/getting-started-kibana

Case de uso da solução ElasticSearch na loja Netshoes

https://www.elastic.co/use-cases/netshoes-pt

**Apache Solr**

O Solr é um pouco mais complexo, mas o resultado é bem similar, podendo ser baixado do endereço https://lucene.apache.org/solr/, seguindo-se os passos de instalação disponíveis em https://lucene.apache.org/solr/guide/8_0/installing-solr.html. Uma vez instalado, é necessário criar um *core* (índice). Para isso, usamos o comando ***bin/solr create_core -c customer*** que vai criar o índice. Ao acessar o endereço http://localhost:8983/solr/#/customer/core-overview, é possível acessar a interface do Solr Admin e já indexar. O Solr também tem o modo Schemaless. Por isso, podemos indexar facilmente sem criar campos. O Solr já vem com um serviço de indexação de arquivos não estruturados, como pdf, doc, xls, dentre outros. Para isso, basta enviar os arquivos para ele. Utilizando o comando post que vem com o Solr, indexei o arquivo pdf do DOU mostrado acima com o seguinte comando **./post -c radar ~/Documentos/pdf/*.pdf**

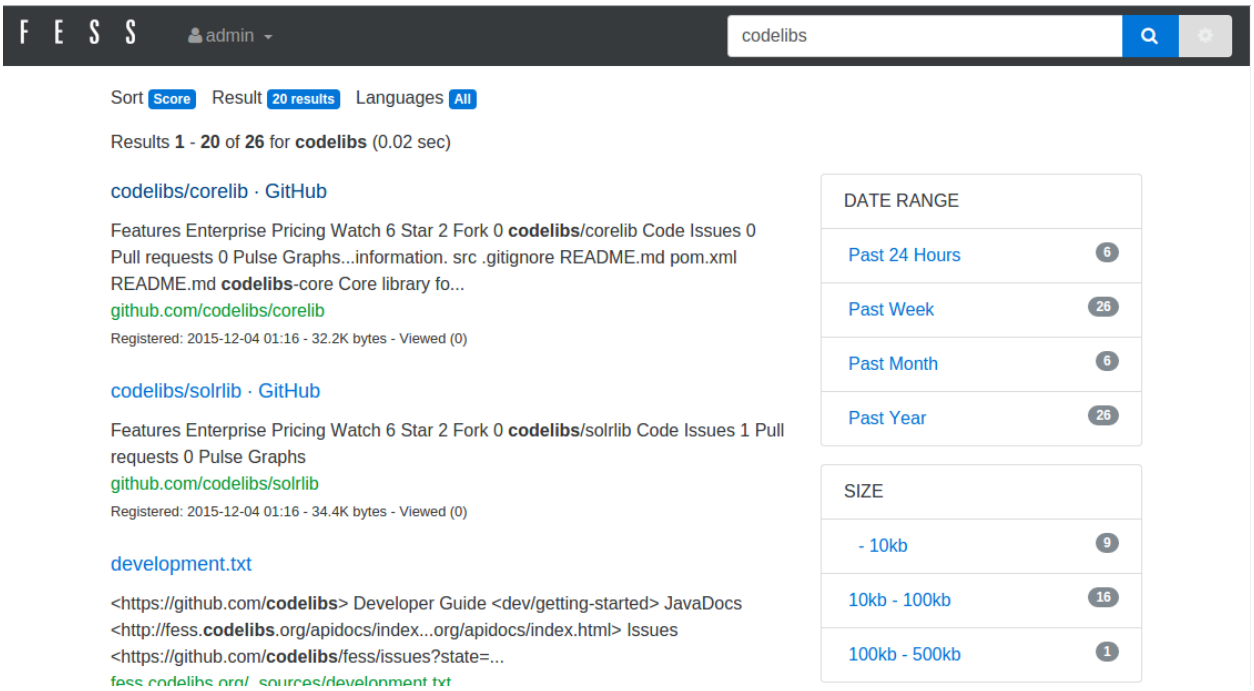
Independentemente de qual será utilizado, boa parte do serviço de indexação e a interface de pesquisa deve ser criada. Outro ponto de atenção é que tanto ElasticSearch quanto Solr não possuem autenticação, devendo ser implementado pela aplicação cliente. As versões comerciais do ElasticSearch possuem o plugin X-Pack com Security. Nas últimas versões passou a disponibilizar-se uma parte básica de segurança. Já no Solr é possível habilitar segurança com o uso do Kerberos.

A distribuição do Hadoop da Cloudera fornece o Solr embutido, denominado Cloudera Search, trabalhando em cluster e armazenando os dados no HDFS. Além de possuir a console Solr Admin para gerenciar o Solr, ele possui a ferramenta Hue, Hadoop User Experience, que permite criar dashboards sobre os dados armazenados no Cloudera Search, além de permitir criar dashboards, dados armazenados no Hive e no Impala.

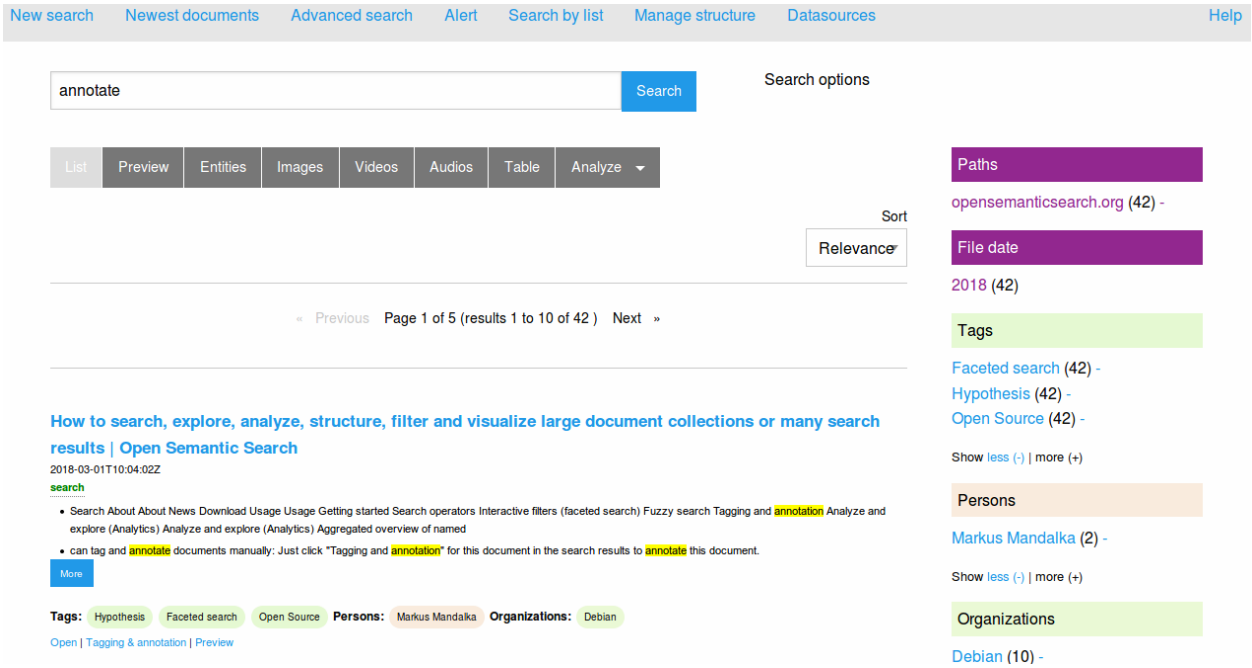
Para saber mais sobre Apache Solr, recomendo a apresentação disponível em https://www.infoq.com/br/presentations/a-revolucao-da-busca-lucene-solr/#mainLogin/

**Interface de busca**

O processo de indexação necessita de um fluxo de processamento para preparar como os dados serão armazenados e pesquisados. Da mesma forma, é necessário criar uma aplicação de pesquisa para que os usuários interajam com as ferramentas de indexação. Muitas vezes, isso é feito pelas próprias aplicações de tramitação de processos. Também podemos utilizar algumas soluções *open source* que fornecem uma interface para facilitar o trabalho. No caso do Elasticsearch, sugiro testar o Free Enterprise Search Server - FESS, disponível em http://fess.codelibs.org, que já vem com ElasticSearch embutido ou pode ser utilizado com um servidor externo, além de disponibilizar um cadastro de usuários e permissões para restringir o acesso aos dados. Para o Solr, sugiro testar o Open Semantic Search, disponível em https://www.opensemanticsearch.org/, que provê um mecanismo de indexação e uma camada de apresentação que permite realizar as pesquisas. Abaixo segue um scheenshot dessas aplicações.

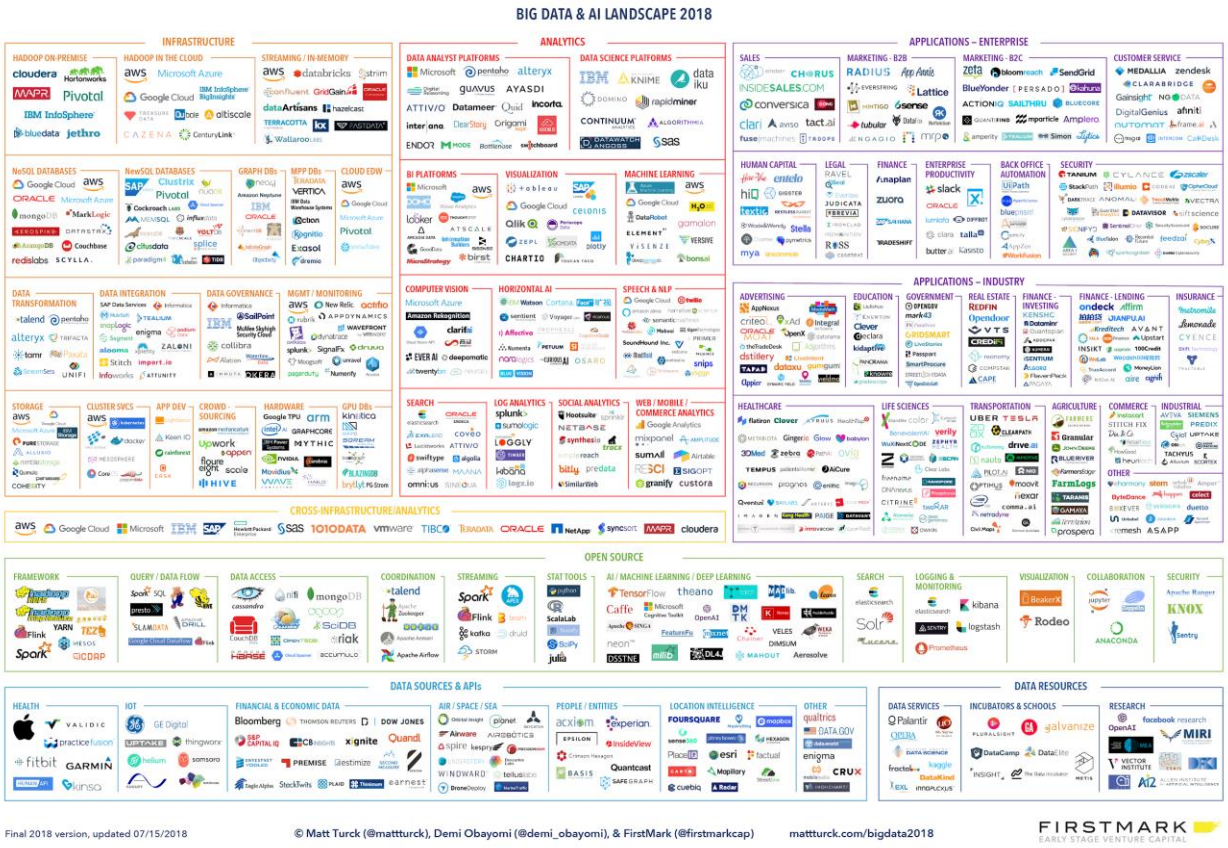



Fonte: http://fess.codelibs.org




Fonte: https://www.opensemanticsearch.org/

O ecossistema de Big Data é muito amplo, com várias ferramentas para o mesmo problema. É importante definir bem o problema a ser resolvido, estudar as várias ferramentas e testar, principalmente, a integração dos vários componentes. Outro ponto fundamental é construir uma solução incremental, ou seja, ir evoluindo aos poucos. Outro ponto fundamental é a proximidade entre a área negocial e a área técnica, para entender a necessidade do negócio, os desafios técnicos e as prováveis soluções. Para se ter uma ideia desse grande universo, a FirstMark publica anualmente o Big Data Landscape com as principais ferramentas, soluções e empresas atuando com Big Data e IA. Abaixo é possível ver uma grande quantidade de ferramentas e empresas, algumas open sources, outras comerciais.




Fonte: https://mattturck.com/bigdata2018/

Links úteis:

https://tika.apache.org/

https://www.xpdfreader.com/pdftotext-man.html

https://www.devmedia.com.br/o-que-e-elasticsearch/40207

https://fess.codelibs.org/

https://www.opensemanticsearch.org/