1 Big Data no Ministério Público

A quantidade de processos que tramitam entre Tribunais e Ministérios Públicos, em todas as instâncias, geram uma grande quantidade de documentos. Segundo dados apresentados pelo CNJ, no relatório Justiça em Números (http://blogs.correiobraziliense.com.br/servidor/cnj-apresenta-justica-em-numeros-2018-com-dados-dos-90-tribunais/), ao final de 2017 eram mais 80 milhões de processos em tramitação. A quantidade de manifestações de juízes, advogados e membros dos MPs é grande, geralmente com dados estruturados e não estruturados. Esse volume de documentos é uma base riquíssima em informações, existem empresas que estão armazenando essas informações em arquitetura de Big Data e aplicando algoritmos de Inteligência Artificial para, por exemplo, descobrir em qual cidade há mais chance de se ingressar com uma ação judicial.

Outra fonte de grandes volumes de dados são as operações de buscas e apreensões. Atualmente, a maioria dos *smartphones* contarem com 64 GB ou mais de capacidade, que são usados para armazenar fotos, vídeos, documentos, *e-mails* e conversas em aplicativos de mensagens. Cada computador pessoal conta com, pelo menos, 1 TB de capacidade. Quando ocorre uma busca e apreensão, é possível imaginar o grande volume de dados recebido. Como exemplo, vamos estimar o volume de dados envolvido em uma operação da Lava Jato amplamente divulgada pela imprensa (https://noticias.r7.com/rio-de-janeiro/balanco-geral-rj/videos/operacao-lava-jato-pf-cumpre-sete-mandados-de-busca-e-apreensao-no-rio-08052019), onde foram realizados 44 mandados de busca e apreensão. Supondo que, para cada alvo da investigação, foram apreendidos um celular e um computador, com os volumes acima citados, teremos um volume aproximadamente de 47 TB de dados.

Outras fontes de dados de grande volume vêm de convênios entre os MPs e outros órgãos para troca de bases de dados. Por exemplo, no Brasil, existem aproximadamente 50 milhões de CNPJs, 250 milhões de CPFs, 160 milhões de eleitores. Para se ter uma ideia do volume, a Receita Federal disponibiliza a base de CNPJ em formato aberto em seu *site*, são aproximadamente 6 GB de dados compactados. Outros dados são

provenientes de quebra de sigilo bancário e telefônico.

É possível perceber que o volume de dados dos MPs formam um verdadeiro *Big* Data.

2 Alternativa aos bancos de dados

Muitas empresas e órgãos públicos armazenam dados, mesmo os não estruturados, em bancos de dados relacionais, muitas vezes, com elevados custos de licenciamento, fazendo uso *storages* de armazenamento de alto desempenho e custo também elevado. Com a implantação de processos eletrônicos, esse problema é agravado, pois todas as peças processuais que antes eram impressas, agora estão armazenadas em sistemas, e as íntegras, na maioria das vezes, são armazenadas em banco de dados.

Para lidar com esse volume de dados, é necessário repensar a arquitetura das aplicações usando soluções de armazenamento e processamento especializados para cada caso de uso. Geralmente, deve-se usar alguma solução de Big Data.

Como coletar, armazenar e analisar dados de diferentes fontes, em vários formatos e gerados em tempos diferentes? As tecnologias que integram o universo de Big Data permite o uso de computadores baratos para formar um *cluster* de alto desempenho com a inclusão de diversas tecnologias para permitir coletar, armazenar e processar dados de acordo com a característica do dado e das análises a serem feitas.

Quando o assunto é banco de dados, temos uma legião de fãs para cada solução comercial e outra legião de fãs para as soluções *open source*. A restrição orçamentária provocada pela Emenda Constitucional 95 tem forçado instituições a avaliar a viabilidade de uso de bancos de dados *open source* como Postgres, MariaDB e MySQL, isso mantendo o paradigma de bancos de dados relacionais. Já os novos requisitos de sistemas têm forçado a adoção de soluções de armazenamento chamadas de bancos de dados NoSQL (Not Only SQL) que possuem soluções para os mais diversos problemas. Esse grupo de soluções é dividido em categorias:

- **Document store**: permite armazenar, em um mesmo documento, diversas

informações referentes a uma entidade. Não exige uma estrutura fixa, dispensando a definição de colunas. Os dados, geralmente, são armazenados por meio de uma estrutura como um json. Assim, pode-se, por exemplo, em um único documento, sob a chave de número do processo, armazenar todos os dados do processo. Geralmente, é mais custoso para gravar algo que geralmente ocorre poucas vezes, mas muito otimizado para a leitura. Principais soluções: CouchDB, ArangoDB e MongoDB (o mais famoso de todos)

- **Graph database**: é muito utilizado em cenários onde o valor está na relação entre as entidades. Assim, é possível percorrer o grafo seguindo as relações. Tem sido muito usada para lidar com problemas relacionados à recomendação e detecção de fraude. Principais soluções: Apache Titan, GraphDB, ArangoDB, e Neo4J (o mais famoso de todos, que possui versão community para uso em apenas um computador. Para uso em *cluster,* requer subscrição)
- **Key-values store**: armazena dados como um conjunto de pares de chave-valor, em que uma chave funciona como um identificador exclusivo. A chave e os valores podem ser qualquer coisa, desde objetos simples até objetos compostos e complexos. É possível armazenar de forma eficiente bilhões de registros, a pesquisa é feita por meio da chave. Principais soluções: MemcacheDB, Redis, DynamoDB
- **Column family store**: os dados são organizados em colunas, e tanto o armazenamento quanto a leitura são baseados nas chaves. HBase e Cassandra.

A adoção dessas soluções não significa abandonar os bancos de dados relacionais. As soluções variam de acordo com o tipo de problema, da quantidade de dados e de clientes usando a solução. Atualmente, as aplicações são híbridas, aproveitando o melhor de cada solução. Como exemplo, podemos ter aplicações utilizando bancos relacionais para armazenar os dados transacionais; bancos de grafos para armazenar dados onde é necessário explorar as relações entre as entidades; e sistemas de busca textual para prover os recursos de indexação e recuperação de informações. Muitas vezes, o armazenando dos dados é feito em HDFS do Hadoop.

3 Data Lake

Data Lake é um termo recente, utilizado para descrever um componente no universo da análise de dados e do Big Data. A ideia é ter um único repositório dentro da empresa, para que todos os dados brutos estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles, geralmente, armazenados no Hadoop.

O Data Lake armazena os dados em seu formato bruto, sem qualquer processamento. Podem ser necessárias várias tecnologias para criar um Data Lake. Apesar de estar totalmente relacionado à Tecnologia, não é um recurso exclusivo da área de TI, é para toda a organização. Todas as partes interessadas devem ser envolvidas no planejamento de projetos de Data Lakes, que serão fundamentais para a arquitetura de Big Data da empresa. Tabelas armazenadas nos bancos de dados relacionais podem ser facilmente replicados para o Data Lake por meio da ferramenta scoop. Para se ter uma ideia do quão simples é para replicar uma tabela de um banco Oracle para o Hive, segue comando para replicar a tabela CREW usando o scoop:

```
sqoop import --connect jdbc:oracle:thin:@localhost:1521/orcl --username
MOVIEDEMO --password welcome1 --table CREW --hive-import
```

Muitas empresas possuem equipes de ciência de dados ou similares nas áreas de negócio, para utilizar os dados do Data Lake e oferecer respostas rápidas para a tomada de decisão.

As soluções de preparação de dados, também conhecidas como ferramentas ETL e as soluções de Business Intelligence (BI) modernas possuem conectores para os mais variados componentes do ecossistema Hadoop, podendo se beneficiar da sua elevada capacidade de processamento. Principais ferramentas analíticas com capacidade de ler dados diretamente do Hadoop por meio de Hive ou Impala:

- MicroStrategy: tradicional ferramenta de BI do mercado. O foco está no uso corporativo, onde é possível criar um Data wharehouse com os dados institucionais. Possui uma versão *desktop* gratuita chamada MicroStrategy Desktop. https://www.microstrategy.com/br

- Power BI: solução de BI da Microsoft criada inicialmente para ampliar as

capacidades analíticas do Excel. Focada no *self-service* BI, ou seja, entregar ao usuário final a capacidade de análise de dados. Possui integração com as Linguagens Python e R. Possui uma versão Desktop gratuita limitada a 1 GB de dados, possui versão em nuvem com licenciamento por usuário ou mesmo para instalação no cliente. https://powerbi.microsoft.com/pt-br/

- QlikSense: ferramenta focada em permitir que usuários finais analisem seus dados. Bastante simples e amigável. Possui uma versão *Desktop* gratuita, limitada à capacidade de memória do computador local. Possui versão server para uso via navegador. https://www.qlik.com/pt-br/products/qlik-sense

Exemplo de conectores do QlikSense: Amazon Redshift, Apache Drill, Apache Hive, Apache Phoenix, Apache Spark, Azure SQL, Cloudera Impala, Google BigQuery, IBM DB2, Microsoft SQL Server, MongoDB, MySQL Enterprise, Oracle, PostgreSQL.

Para ampliar o conhecimento, indico a apresentação disponível em https://www.infoq.com/br/presentations/a-jornada-para-implementacao-de-um-data-lake/

4 Profissionais

Para atuar com Big Data é necessário ter uma equipe multidisciplinar para montar e sustentar a solução. Um dos maiores desafios para o serviço público é ter no seu corpo técnico profissionais com tais conhecimentos. Por isso, muitas vezes, as instituições contratam empresas terceirizadas. Os principais perfis profissionais são:

- Engenheiro de Dados: responsável por garantir que os dados estarão disponíveis para a análise de forma segura. Montam a infraestrutura para armazenamento e

processamento de grandes conjuntos de dados não é tarefa fácil. Hadoop, Spark, Cassandra, Hive, Hbase, Pig, Sqoop, MongoDB, API de integração. São os responsáveis pela criação do Data Lake.

- Cientista de Dados: são mineradores de dados. Recebem uma enorme massa de dados desorganizados (estruturados, semiestruturados ou não-estruturados) e usam suas habilidades para limpar, tratar, transformar e organizar esses dados. Em seguida, aplicam suas capacidades analíticas para descobrir soluções para os problemas de negócios e contribuir na tomada de decisões e estratégias empresariais.

As equipes, geralmente, são multidisciplinares, com profissionais de várias formações: estatísticos, matemáticos, advogados, além, é claro, dos profissionais de TI.

5 Aplicações

As soluções de Big Data, muitas vezes, associadas com Inteligência Artificial, oferecem recurso para a construção de aplicações para:

- Armazenamento de grandes volumes de dados
- Processamento de grandes volumes de dados
- Soluções analíticas
- Mineração de textos
- Mineração de dados
- Processamento de Linguagem Natural
- Transcrição de áudios
- Reconhecimento facial e de objetos
- Reconhecimento de caracteres
- Sistema de buscas
- Detecção de fraudes

5 Como testar Hadoop

Uma forma de testar essas tecnologias é utilizando aplicações em nuvem, pois facilitam a instalação de cada componente. Os principais provedores Amazon, Microsoft

e Google oferecem esses componentes ou componentes similares. Muitas instituições possuem restrição de uso de computação em nuvem, podendo instalar localmente em um computador único essas soluções.

Os principais fornecedores de Hadoop como Cloudera e Hortonworks, esta última adquirida pela primeira em 2019, fornecem imagens de máquinas virtuais ou docker para teste. Recomenda-se instalar o VirtualBox para permitir criar máquinas virtuais e baixar a respectiva versão. Segue *link* para baixar e testar a versão da Cloudera do Hadoop.

https://www.cloudera.com/downloads/quickstart_vms/5-13.html

Segue vídeo mostrando como instalar o Cloudera em uma máquina virtual https://www.youtube.com/watch?v=HP4g2BU7-xU

Links Úteis:

https://aws.amazon.com/pt/nosql/

http://www.cienciaedados.com/data-lake-a-fonte-do-big-data/

http://datascienceacademy.com.br/blog/os-4-estagios-para-construir-um-data-lake-de-forma-eficiente/

https://medium.com/data-hackers/o-guia-semi-definitivo-para-data-lakes-461b1878697f

http://datascienceacademy.com.br/blog/10-carreiras-em-big-data-e-data-science/

https://blog.oncase.com.br/cloudera-com-kerberos-pentaho-do-zero/