FACULDADE METROPOLITANA DO ESTADO DE SÃO PAULO - FAMEESP

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

RELATÓRIO DO CICLO DE VIDA PARA O PROCESSO DE CIÊNCIA DE DADOS.

ALUNA: Rosilene Lima Justo

Brasília – DF
01/10/2024

RELATÓRIO DO CICLO DE VIDA PARA O PROCESSO DE CIÊNCIA DE DADOS.

Data de entrega do relatório: 01/10/2024

Responsável:
Relatório: Rosilene Lima Justo – Cientista de dados

Versão 1.0

1. Introdução

O hospital *Newton Turing* é referência em atendimento com mais de 1.500 pacientes com diversas especialidades e possui funcionamento 24h. O diretor insatisfeito com a falta de precisão na tomada de decisão e a falta de agilidade nos procedimentos, optou por contratar um especialista em ciência de dados para a obtenção de dados estatísticos em relação aos pacientes e análises preditivas para se preparar em relação à evolução das doenças, a contratação de mais funcionários e à aquisição de materiais médicos e medicamentos.

O presente relatório tem como objetivo relatar a análise do ciclo de vida para o processo de implantação de ciência de dados no hospital *Newton Turing*. De acordo com os estudos realizados, o ciclo de vida para o processo de ciência de dados possui fases: coleta, armazenamento, recuperação e descarte; fatores: privacidade, integração, qualidade, direitos autorais, disseminação e preservação, que intervêm em todas as fases. Possui metodologias importantes para o seu desenvolvimento, assim, como também ferramentas para realizar as análises.

2. Ciclo de Vida para o processo de ciência de dados

O ciclo de vida para o processo de ciência de dados é composto pelos passos necessários para a extração do conhecimento daquele conjunto de dados. Segundo Sant'Ana (2016), o ciclo de vida dos dados é composto por quatro fases: coleta, armazenamento, recuperação e descarte. Além disso, há alguns fatores que permeiam todas as fases: privacidade, integração, qualidade, direitos autorais, disseminação e preservação. Veremos cada uma delas e as suas respectivas ferramentas, assim como, uma metodologia usada para agilizar o trabalho das equipes.

2.1 Coleta

O passo de coleta ocorre quando se identifica o conjunto de dados que representa o objetivo de estudo.

Ferramentas: Pentaho, Sqoop, Spark, Streaming, Kafka e Flume.

Fatores:

Qualidade: É a medida da integridade dos dados em diversas dimensões, como: precisão, integridade, consistência, confiabilidade e etc. O gerenciamento da qualidade dos dados consiste em: planejamento, implementação e controle das atividades para mensurar, melhorar, e garantir a conveniência dos dados para o uso. O DMBOK um guia sobre gerenciamento de dados, governança, e qualidade dos dados sugere-se que haja um responsável que coleta e armazena os dados, que deve responder pela veracidade desses dados e pelo quanto eles representam a realidade. Também criação de métricas de qualidades de dados, de modo a planejar como mensurá-los e melhorá-los.

Privacidade: É o princípio que garante o controle de como e onde os dados são coletados. A Lei de n° 13.709/2018, Lei Geral de Proteção de Dados (LGPD) é uma lei que tem como objetivo proteger os direitos de privacidade das pessoas. A base da lei é o consentimento ou autorização do titular dos dados, antes do tratamento ser realizado.

Disseminação: Pode ser entendida como propagação, ou seja, a informação sendo difundida por vários meios e suportes abrangendo um determinado usuário com base no esquema tradicional de emissor, canal, mensagem, receptor.

Direitos autorais: A Lei de n° 9.610/1998 não protegem dados, fatos ou informações, mas sim à forma como eles são expressos e comunicados. A violação dos direitos autorais é considerada crime e está prevista no artigo 184 do Código Penal. Os dados privados têm uma titularidade, e o seu uso deve ser observado.

Preservação: A preservação dos dados coletados é necessária para identificar o problema e a forma como serão analisados.

Integração: A integração de dados é um processo que reúne dados de diferentes fontes, formatos e estruturas, para que eles possam ser analisados e usados de forma unificada.

Procedimentos a serem realizados no ambiente:

Com a equipe estruturada foi observado a necessidade de fazer um levantamento com entrevistas na intenção de direcionar a necessidade de todos os departamentos do hospital. Também foi necessário remanejar uma pessoa da equipe para organizar os dados que estão em prontuários e fichas e passar para planilhas obtendo arquivos mais precisos. De acordo com o art. 6° ao 14° da LGPD, é necessário ter autorização dos usuários em formulários para o armazenamento dos dados. Nos formulários deverá ter explicitado qual será o uso e finalidade dos dados. Sempre informar ao usuário. Se houver alguma violação será necessário notificar às autoridades competentes. Será preciso fazer unificação dos dados em um mesmo ambiente para que seja possível à análise. Uma vez identificado os dados que se deseja analisar o próximo passo é o armazenamento.

2.2 Armazenamento

O passo do armazenamento é o processo de guardar, manter e preservar informações digitais para o uso futuro ou presente.

Ferramentas: Pode ser usados os Banco de dados relacionais como *PostgreSQL* e *MySQL*, bem como o *Apache Hadoop*, para o *Big Data*.

Fatores:

Qualidade: É um fator importante para o armazenamento de dados, pois indica o nível de conformidade dos dados com um padrão de regras.

Privacidade: A LGPD estabelece diretrizes para a proteção de dados pessoais. O armazenamento e processamento de dados pessoais só podem ser feitos com o consentimento explícito do usuário.

Disseminação: A disseminação é analisada segundo a perspectiva da transferência de informação em face à reconfiguração da ideia de cidadão e cidadania.

Direitos Autorais: A Lei 9.610/98 protege aas obras intelectuais, literárias ou científicas. Os dados primários da biodiversidade são considerados descobertas (inéditos) e, portanto, integram a propriedade intelectual e são protegidos como tal. Para a LGPD de forma simples, a transferência internacional de dados pessoais ocorre sempre que algum dado pessoal (referente a uma pessoa física, via de regra) é enviado, armazenado ou acessado de um país estrangeiro. No art. 16 da LGPD é vedado o compartilhamento com terceiros dos dados obtidos fora das hipóteses permitidas em Lei.

Preservação: A preservação de dados no armazenamento é importante para manter a segurança e a eficiência do acesso e gerenciamento das informações digitais. Existe atualmente algumas técnicas para preservação de grandes volumes de dados.

Tipos integração: Integração em lote, integração em tempo real, consolidação de dados, virtualização de dados.

Procedimentos a serem realizados no ambiente:

Junto com a equipe é necessário determinar quais ferramentas e softwares que deveram ser usados para o armazenamento. O processo de armazenamento de dados é feito pela solicitação, transferência de dados, gravação, confirmação de gravação. Existe dois tipos de armazenamento o primário e o secundário. Os principais dispositivos para o armazenamento de dados corporativos é: *Direct Attached Storage* (DAS); Sistemas de armazenamento conectados à rede (NAS); Redes Armazenamento Corporativas (SAN); *Object storages*; *Cloud computing*. Os principais protocolos de comunicação usados em sistemas de armazenamento é: *SATA; SCSI; SAS; ICSI; NFS; CIFS/SMB; Fibre Channel; HTTP/HTTPS.*

As vantagens de escolher um sistema de armazenamento eficiente é:

1-Desempenho aprimorado;
2-Redução de custos;
3-Melhor utilização do espaço;
4-Resiliência aprimorada;
5-Gerenciamento simplificado;
6-Sustentabilidade.

Como foi identificado que os dados são volumosos será preciso uma estrutura específica de Big Data. Pode ser utilizado o *Apache Hadoop.* Um serviço que contém o sistema de arquivos distribuídos HDFS, que proporciona segurança para o armazenamento e posterior recuperação dos dados para o processamento. Também pode ser utilizados outros serviços do ecossistema Hadoop, tais como:

Ambari – Gerenciamento, provisionamento e monitoramento dos clusters do Hadoop;
Zookeeper – Coordenação;
Dozie – Workflow;
Pig – Scripting;
Mahout – Machine learning;
Hive – Data Warehouse;
Map Reduce – Processamento de dados;
HDFS – Sistema de arquivos distribuídos;
Hbase – Banco de dados NoSQL;

Sqoop / REST / DDBC – integração de dados.

2.3 Recuperação

É o processo de restaurar informações a fim de que se possa realizar os processamentos iniciais para a exploração dos dados e, posteriormente, para a extração de conhecimento.

Ferramentas: SGBDs com o PGAdmin, no caso do PostgreSQL, O Apache Drill e o Apache Hive.

Fatores:

Qualidade: A garantia de qualidade de recuperação de dados (QA) é o processo de assegurar que os dados recuperados de um hardware sejam precisos, completos e utilizáveis. Para isso, são verificadas a integridade, a legibilidade e a funcionalidade dos dados.

Privacidade: A LGPD é aplicável à recuperação judicial e deve ser observada pelo administrador judicial, que deve agir com prudência e respeitar os princípios da necessidade, da finalidade e da adequação. A LGPD é aplicável a qualquer operação de tratamentos de dados pessoais.

Disseminação: Diante da dinâmica de informacionalização da internet, cujo volume de conteúdos gerados não tem precedentes, certamente a disseminação e a recuperação das informações tornam-se mais difíceis e imprecisas; neste cenário em que o pesquisador não consegue acompanhar o volume de informação disponibilizado, a busca e recuperação da informação é um desafio

Direitos Autorais: Os direitos autorais não protegem fatos, informações ou dados, e nem mesmo o conteúdo de uma obra, pois seu objeto de proteção é a forma literária ou artística em que são expressos, comunicados. No entanto, quando estes elementos são agrupados, organizados ou sistematizados em uma base ou banco de dados que seja minimamente original com relação à seleção, organização ou disposição de seu conteúdo, este material passa a ter seu acesso e utilização controlado pelo titular deste conjunto, então protegido por direitos autorais

Preservação: É necessário garantir preservação e rápida recuperação em casos de eventualidades, uma vez que estes dados contêm informações cruciais para o bom funcionamento dos negócios das empresas.

Integração: É necessário usar ferramentas eficientes para a recuperação e a integração dos dados.

Procedimentos a serem realizados no ambiente:

Com base no que se deseja analisar, é definido o método de recuperação, que pode ser a restauração a partir de backups ou recuperação de sistemas de arquivos. Os sistemas de recuperação da informação podem ser definidos como um conjunto de operações consecutivas

executadas para localizar, dentro da totalidade de informações disponíveis, aquelas realmente relevantes. Para isso, executam as funções de seleção, análise, indexação e busca das informações.

É necessário que a equipe faça as análises utilizando uma linguagem de programação dentro do Apache Hadoop. É sugerido a linguagem R para unificar as fontes com os dados, realizar uma análise exploratória e realizar as devidas estatísticas.

2.3 Descarte

Após a recuperação dos dados, é possível verificar que há dados que podem ser descartados, pois não fazem sentido ao problema, têm baixa veracidade ou, ainda, contêm campos nulos ou diversos erros.

Ferramentas: Cada ferramentas de armazenamento tem operações específicas para o descarte dos dados. Algumas aceitam comandos em SQL e outras em seus próprios comandos. Também pode ser desenvolvido procedimentos para descarte automático, baseado em linguagens de programação, como Python, R e Java.

Fatores:

Qualidade: A LGPD estabelece que os dados pessoais serão eliminados após o término do seu tratamento, autorizada a conservação para a finalidade de cumprimento de obrigação legal ou regulatória, de estudo por órgão de pesquisa, de transferência a terceiro ou de uso exclusivo pelo controlador, prezando sempre pela anonimização quando possível.

Privacidade: A LGPD estabelece que alguns documentos originais não podem ser eliminados após a digitalização, como documentos de identificação, documentos de porte obrigatório e documentos referentes a operações e transações financeiras. O descarte de dados pessoais dos titulares deve ocorrer de forma adequada, mesmo quando constantes em meios físicos.

Disseminação: Os arquivos digitais têm o mesmo valor legal, e a LGPD estabelece que os dados pessoais devem ser eliminados após o término do seu tratamento, exceto quando necessário para cumprir obrigações legais ou regulatórias, para fins de pesquisa, transferência a terceiros ou uso exclusivo pelo controlador, com a devida anonimização quando possível.

Direitos Autorais: Importante frisar, ainda, o artigo 15º da LGPD, o qual estipula que a cessação do processamento de dados pessoais deve ocorrer assim que a finalidade original for atingida, ou quando os dados já não são mais necessários ou pertinentes para essa finalidade.

Preservação: Art. 37 da LGPD O controlador e o operador devem manter registro das operações de tratamento de dados pessoais que realizarem, especialmente quando baseado no legítimo interesse. De acordo com o art. 40 da LGPD a autoridade nacional poderá dispor sobre padrões de interoperabilidade para fins de portabilidade, livre acesso aos dados e segurança, assim como sobre o tempo de guarda dos registros, tendo em vista especialmente a necessidade e a transparência.

Integração: Conforme a LGPD é necessário o descarte dos dados solicitados de exclusão de todas as fontes. A ação de descarte de dados deverá ser registrada, para que se mantenha a rastreabilidade dessa atividade.

Procedimentos a serem realizados no ambiente:

Todos os dados existentes, físicos ou virtuais, devem ser devidamente destruídos quando uma solicitação nesse sentido é apontada. Para além disso, a mídia física deve ser descartada de maneira adequada após o uso, o que exige práticas cabíveis, como a trituração ou a desmagnetização. A LGPD estabelece que os dados pessoais devem ser mantidos apenas pelo tempo necessário para a finalidade a que se destinam. Por isso, é importante adotar medidas para garantir a sua exclusão após o término do prazo estabelecido. A reciclagem é uma opção para os resíduos gerados pelo descarte de documentos, desde que seja feita de acordo com os critérios ambientais.

Após realizar a análise exploratória é descartado todos os dados inválidos para análise, é necessário que tudo seja registrado e formado metadados para explorações futuras. Com as métricas estabelecidas para exploração do que deseja de insights dos dados é preciso desenvolver visualizações em formas de gráficos. Alguns gráficos como:

Pizza – para valores em porcentagem com até 4 dimensões.

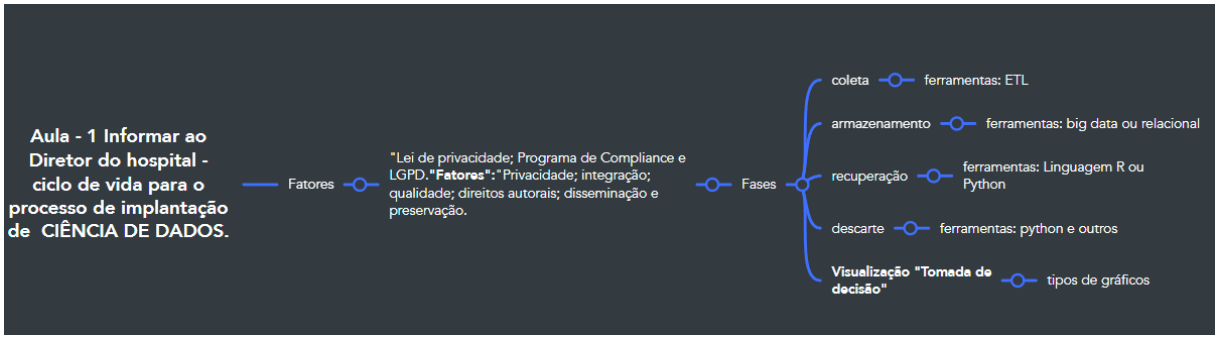
Gráfico de Barra – Para insights em anos, meses ...

Gráfico de colunas – Para insights em anos e setores na forma Vertical.

Gráfico de linhas e áreas – mostra sequência númerica.

Gráfico em rede – medição de termos específicos estatísticos.

Figura – 1 Mapa mental do ciclo de vida do processo de ciência de dados.



Fonte: Rosilene Lima Justo, 2024

A Microsoft desenvolveu o TDSP, que é uma metodologia de ciência de dados ágil e interativa. O TDSP incorpora as práticas recomendadas e as estruturas da Microsoft e de outros líderes do setor para ajudar as equipes a implementar iniciativas de ciência de dados com eficiência.