2023-2024



Projeto de processamento estruturado de infomração

Trabalho Prático – época normal

**Trabalho elaborado por:**

**Grupo 19**

8220169 – César Ricardo Barbosa Castelo

8220337 – Hugo Ricardo Almeida Guimarães

8220307 – Pedro Marcelo Santos Pinho

Índice

[Índice de Figuras 2](#_Toc156377095)

[Chave de Siglas 3](#_Toc156377096)

[Introdução 4](#_Toc156377097)

[Requisitos que limitam o domínio 5](#_Toc156377098)

[1. Mongo DB 6](#_Toc156377099)

[a. Estrutura da base de dados 6](#_Toc156377100)

[a. Clientes 6](#_Toc156377101)

[b. Produtos 7](#_Toc156377102)

[c. Vendas 7](#_Toc156377103)

[d. Devoluções 8](#_Toc156377104)

[b. Importação dos dados fornecidos 9](#_Toc156377105)

[c. Integração do BaseX com o MongoDB 11](#_Toc156377106)

[2. Organização do XML 13](#_Toc156377107)

[a. Regras (XML Schema) 13](#_Toc156377108)

[MongoDB Charts 14](#_Toc156377109)

[Apreciação Crítica (acabar o primeiro parágrafo) 15](#_Toc156377110)

# Índice de Figuras

[Figura 1 - coleções do mongodb 6](#_Toc156377148)

[Figura 2 - Estrutura da coleção cliente no mongoDB 6](#_Toc156377149)

[Figura 3 - Estrutura da coleção produtos no mongoDB 7](#_Toc156377150)

[Figura 4 - estrutura da coleção vendas no mongodb 7](#_Toc156377151)

[Figura 5 - Estrutura da coleção devoluções no mongodb 8](#_Toc156377152)

[Figura 6 - script inicial para criação da base de dados e coleções 9](#_Toc156377153)

[Figura 7 - script para a criiação de índices 9](#_Toc156377154)

[Figura 8 - pipelines para a transformação dos dados 10](#_Toc156377155)

[Figura 9 - Relatório de vendas devolvido pelo baseX 11](#_Toc156377156)

[Figura 10 - Relatório de devoluções devolvido pelo baseX 12](#_Toc156377157)

[Figura 11 - Ficheiros com as regras para montar as regras dos relatórios 13](#_Toc156377158)

[Figura 12 - Regras dos relatórios 13](#_Toc156377159)

[Figura 13 - Gráficos criados no mongoDB charts 14](#_Toc156377160)

# Chave de Siglas

|  |  |
| --- | --- |
| XML | Extensible Markup Language |
| XSD | XML Schema Definition |
| CSV | Comma Separated Values |
| JSON | Java Script Object Notation |

# Introdução

Este trabalho foi realizado para o âmbito da disciplina de *Processamento Estruturado de Informação*, que funcionará como integrador dos conhecimentos adquiridos no decorrer das aulas.

Este trabalho consiste na construção de um vocabulário XML para suportar o envio de relatórios de vendas e devoluções de cada parceiro da empresa *Phone for You*, uma empresa que comercializa smartphones através de várias lojas. Assim cada parceiro terá de implementar nos seus sistemas informáticos, um módulo que suporta a geração de documentos XML de vendas e devoluções de acordo com vocabulário estabelecido.

Para além da construção de um vocabulário XML para suportar o envio de relatórios de vendas e devoluções, também será preciso fazer uma transformação dos dados já existentes de vários ficheiros CSV para uma base de dados orientada a documentos (MongoDB), para além disso, também será necessário fazer uma API para aceder aos dados, e construir os relatórios para depois retornar ao utilizador.

# Abordagem

Após a leitura do enunciado

# Requisitos que limitam o domínio

Ao desenvolver o relatório, foram considerados requisitos específicos que impuseram limitações ao domínio dos elementos e atributos utilizados no vocabulário. Essas restrições foram estabelecidas para garantir a coerência, padronização e integração adequada com outros sistemas, sendo elas:

# Mongo DB

## Estrutura da base de dados

Para a realização deste trabalho, foi-nos enviado vários ficheiros, em formato CSV, que representam um subconjunto de informação tipicamente armazenada por um dos parceiros da *Phone for You*, com o objetivo de fazer a importação deles para o Mongo DB. No entanto, como a informação de ficheiros estava muito fragmentada, uma importação direta era inviável, já que as consultas iriam demorar imenso tempo. Então para resolver esse problema, fez-se uma transformação dos dados quando estes foram importados. Os dados ficaram divididos segundo as seguintes coleções:

Gráfico

Descrição gerada automaticamente com confiança baixa

Figura - coleções do mongodb

### Clientes



Figura - Estrutura da coleção cliente no mongoDB

Para poder importar informação dos clientes, foi preciso juntar vários ficheiros em uma única coleção, sendo esses ficheiros: “*address.csv”*, “city.csv” e “country.csv”.

Como um cliente tem uma morada registada, e cada morada tem de ter registado uma cidade, e uma cidade está registada a um país, decidiu-se juntar a informação dessas tabelas em uma única coleção.

### Produtos

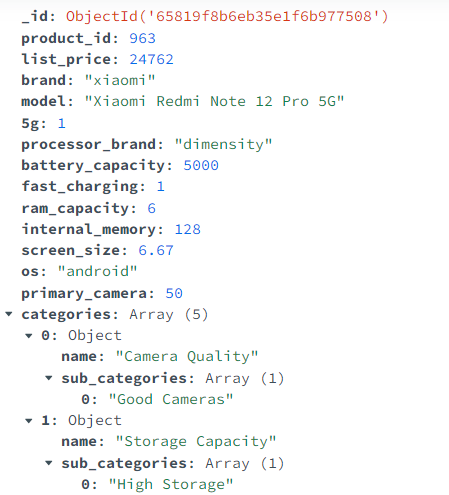


Figura - Estrutura da coleção produtos no mongoDB

Para os produtos aconteceu algo semelhante à coleção dos clientes, no entanto em vez de se juntar a morada, juntou-se o produto com as categorias, visto que um produto tem uma ou mias categorias, e uma categoria pode ter uma ou mais subcategorias. Juntando assim a informação dos ficheiros *“Product.csv”*, *“sub\_category\_product.csv”*, *“sub\_category.csv”* e *“category.csv”*.

### Vendas



Figura - estrutura da coleção vendas no mongodb

Para poder criar a coleção das vendas, foi preciso unir os ficheiros: *“sales\_header.csv”*, *“sales\_lines.csv”*, pois, uma “sales\_header” pode ter uma ou mais “sales\_lines”. Para além de unir esses dois ficheiros, também se colocou alguma informação do cliente que comprou o produto, e dos produtos comprados, assim evita-se ter de percorrer várias coleções para quando se for criar relatórios, fazendo com a aplicação tenha um desempenho muito melhor.

### Devoluções

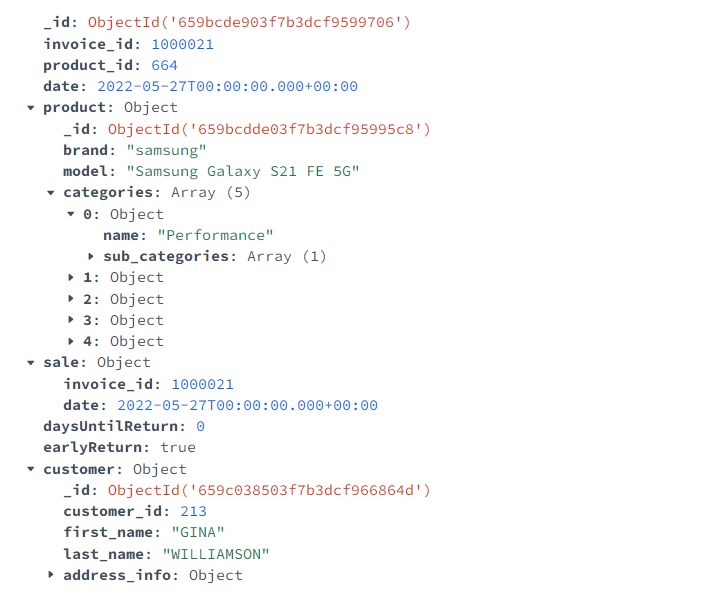


Figura - Estrutura da coleção devoluções no mongodb

Nas devoluções, juntou-se um pouco de toda a informação das outras coleções criadas anteriormente, pois, uma devolução é efetuada por um cliente, sobre um produto em específico, e esse produto está associado a uma venda a partir do código da sua fatura (invoice\_id).

Uma característica do MongoDB, em comparação com outros modelos de bases de dados, é a sua abordagem semiestruturada. Esta abordagem confere flexibilidade ao MongoDB, que se aproveita da redundância para proporcionar um alto desenpenho. Por exemplo, na coleção de clientes, cada cliente possui informações de endereço armazenadas em "address\_info". Notavelmente, o mesmo cliente também terá informações à cerca do endereço registadas nas devoluções. Caso o cliente mude de endereço, a alteração será efetuada apenas na coleção clientes, sem afetar as informações nas devoluções. Este design flexível e eficiente destaca a capacidade do MongoDB de lidar com dados semiestruturados de maneira ágil e eficaz, por isso tentou-se ao máximo durante as migrações dos dados abusar da redundância, para que as consultas conseguissem ser feitas de forma quase instantânea.

## Importação dos dados fornecidos

Para importar os dados fornecidos para a base de dados, primeiramente, foi preciso criar um script que criasse uma uma base de dados vazia no MongoDB, e depois criasse várias coleções, que irão conter os dados dos ficheiros csv, esse ficheiro chama-se *“setting up the database.txt”*, e ele está na pasta *“.\MongoDB”*

Uma imagem com texto, captura de ecrã, Tipo de letra, file

Descrição gerada automaticamente

Figura - script inicial para criação da base de dados e coleções

Nota: o ficheiro apenas cria coleções vazias, para fazer a importação dos dados dos ficheiros, é preciso fazê-la manualmente a partir do MongoDB Atlas.

Após ter todas as coleções com os dados originais importados, vai ser preciso executar o script de outro ficheiro chamado *“Setting indexes after csv inport.txt”,* ele também esá localizado na pasta *“.\MongoDB”, e tal como o nome indica, ele vai defenir os índices necessários para as migrações que se vai fazer, em alguns casos, as migrações nem são possíveis de realizar sem estes ínidces*



Figura - script para a criiação de índices

Com os dados importados e os índices criados, vai ser preciso fazer uma migração desses dados para novas coleções, que conterão os dados organizados de uma forma que o MongoDB consiga ler de forma eficiente. Para isso, dentro da pasta*”.\MongoDB\pipelines”* contém 4 scripts que contêm todas contêm todas as pipelines de agregação[[1]](#footnote-1) necessárias para a migração dos dados. Elas estão todas numeradas de 1 a 4 para dizer que elas devem ser executadas por ordem crescente, dessa forma, permite que os últimos ficheiros, que contêm mais informação e necessitam de mais alterações, não precisassem de fazer tantas transformações, bastava aceder às coleções que já têm os dados prontos para poder fazer a transferência, poupando assim bastante tempo e esforço.

Uma imagem com texto, captura de ecrã, Tipo de letra, algebra

Descrição gerada automaticamente

Figura - pipelines para a transformação dos dados

## Integração do BaseX com o MongoDB

Para a integração do baseX com o mongoDB, foram desenvolvidas duas consultas na data Api do mongoDB, para receber relatórios de vendas e devoluções para um determinado mês. Para poder acededer a essas consultas, foi desenvolvida uma API no baseX que comunica com a data API do mongoDB, que devolve os dados e transforma-os para o formato xml. É graças a esta API que é possível especificar o mês e o ano dos dados que queremos receber, pois é nela que eles são definidos como parâmetros.

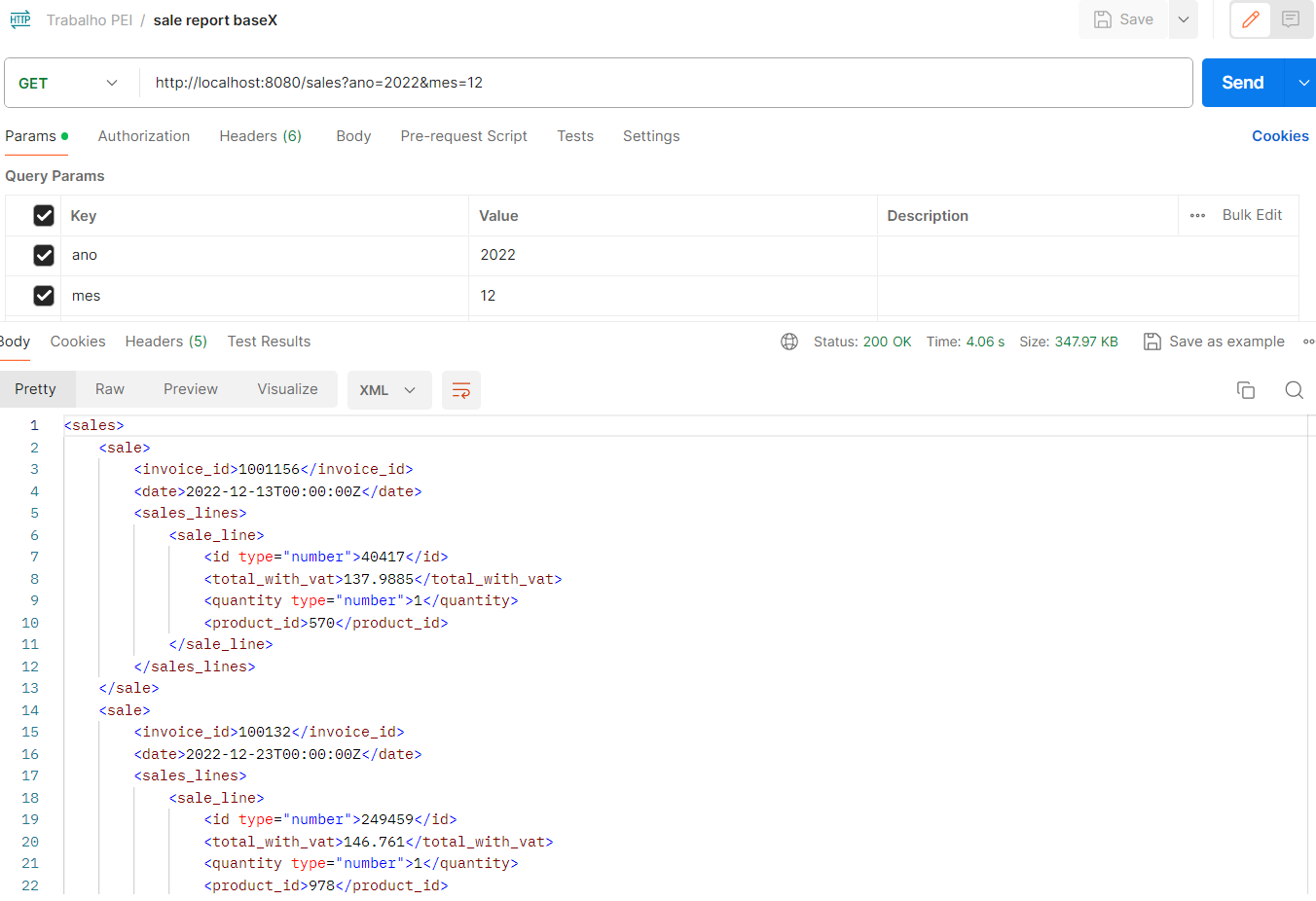


Figura - Relatório de vendas devolvido pelo baseX

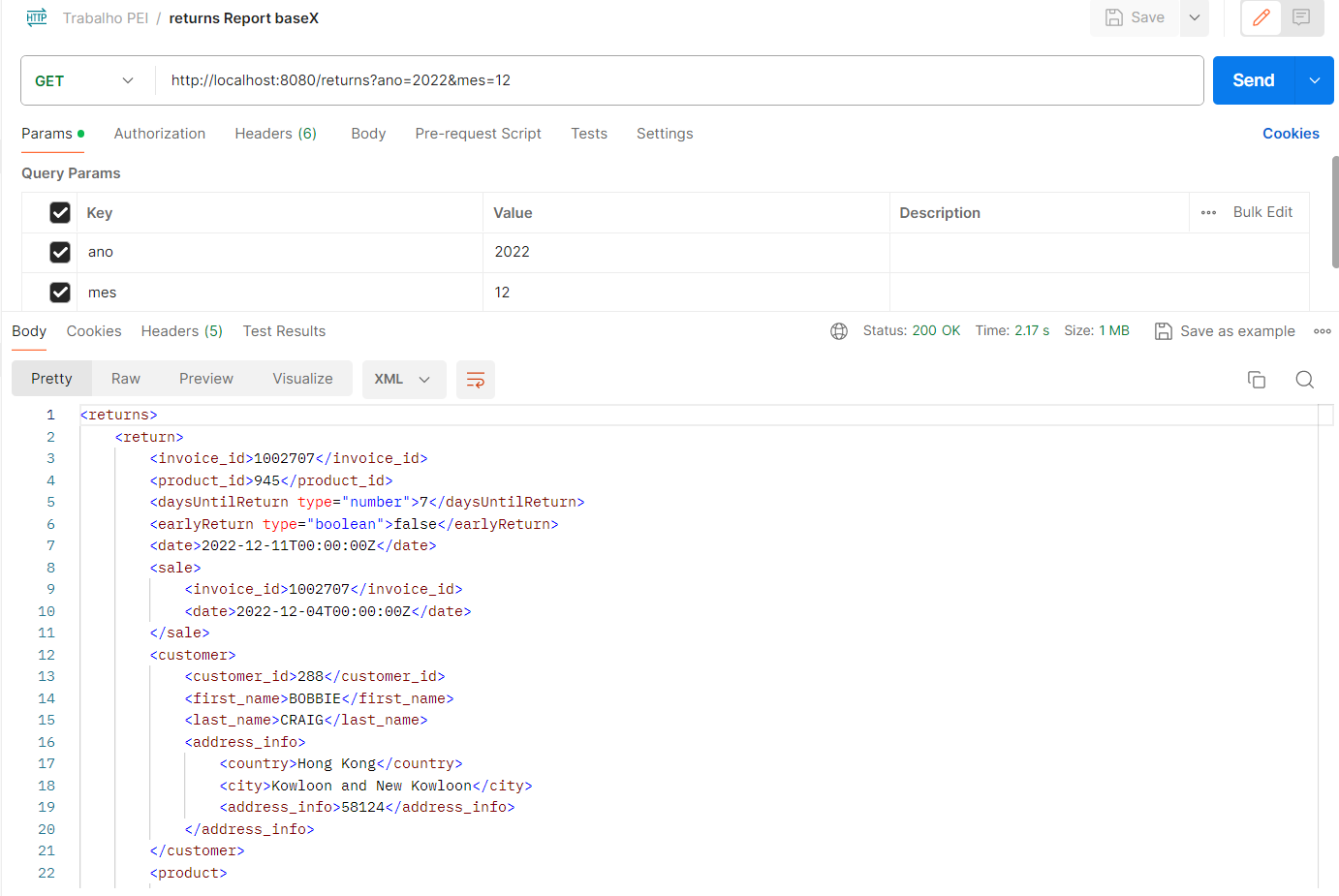


Figura - Relatório de devoluções devolvido pelo baseX

# Organização do XML

## Regras (XML Schema)

Para as regras do XMl, tentou-se ao máximo com que elas fossem o mais reutilizáveis possíveis, então criou-se vários ficheiros, onde cada um deles define regras específicas para um componente, como pode ser vito na figura a seguir:

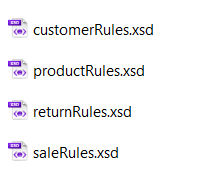


Figura - Ficheiros com as regras para montar as regras dos relatórios

Estas regras encontram-se na pasta *“.\BaseX\xsd\rules”*.

Por estarem em ficheiros diferentes, possibilita com que seja possível montar um relatório costumizado, como por exemplo: um relatório de vendas; devoluções; vendas e devoluções; etc..

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Figura - Regras dos relatórios

Os ficheiros mostrados na figura de cima, são os ficheiros com as regras usadas na API do baseX para validar os resultados obtidos. Estes ficheiros encontram-se na pasta *“.\BaseX\xsd”*.

# MongoDB Charts

Como o MongoDB fornece uma interface intuitiva e personalizável, permitindo a criação rápida de gráficos interativos, decidiu-se usá-la neste projeto para criar gráficos a partir dos valores lá armazenados. Desta forma podemos ter um guia visual, de como uma informação se relaciona com a outras, permitindo uma maior annálise sobre o negócio, esta é uma estratégia muito usada por lojas e empresas para identificarem e superarem adversidades. Em última análise, a escolha do MongoDB Charts representa uma abordagem eficaz para maximizar a utilidade dos dados disponíveis e impulsionar o sucesso de um projeto

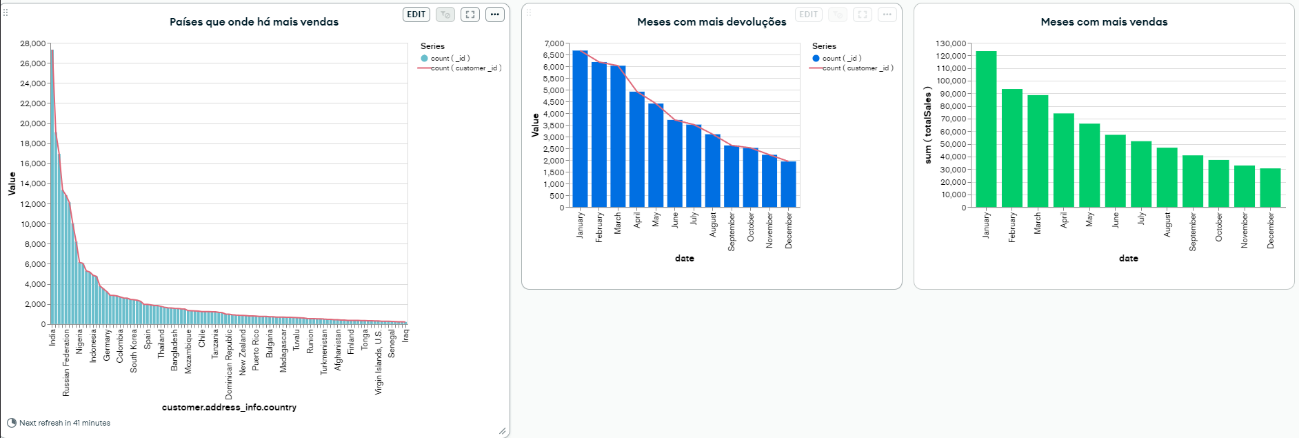


Figura - Gráficos criados no mongoDB charts

**Link para poder aceder aos gráficos:** <https://charts.mongodb.com/charts-project-0-voykk/public/dashboards/659fddb5-26de-44b5-821b-79606e9900af>

# Apreciação Crítica (acabar o primeiro parágrafo)

Ao longo deste relatório, foi moostrado como foi realizada a migração dos dados em texto para uma base de dados orientada a ficheiros, como foram feitas as consultas para ficarem de acordo com o vocabulário XML também desenvolvido neste projeto.No entanto, mesmo após a conclusão do trabalho, podem ser identificadas algumas falhas que não puderam ser corrigidas atempadamente, sendo elas: a identificação do parceiro que gerou o relatório de vendas e devoluções e a inserção de alguns campos adicionais pedidos

Ao longo da realização do projeto deparamo-nos com vários desafios, como por exemplo, a parte da criação da migração dos dados dos ficheiros CSV para o mongoDB, a manipulação de datas e horas, e a passagem dos relatórios de JSON para XML a partir do baseX, que de acordo com todos os elementos do grupo, foi o que deu mais trabalho. Infelizmente, alguns dos desafios encontrados não conseguiram ser ultrapassados, como:

Mas, mesmo com todos esses desafios, podemos concluir que a realização deste trabalho foi bem condebia tendo em conta o que nos foi proposto e também as nossas próprias exigências para o trabalho, entre elas, a constante busca pela estrutura da base de dados, que conseguisse gerar relatórios da forma mais eficiente possível. Este projeto mostrou-se muito importante para o nosso desenvolvimento, pois conseguimos colocar em prática aquilo que foi lecionado durante as aulas, assim, sedimentando os nossos conhecimentos.

1. Um pipeline de agregação consiste em um ou mais estágios que processam documentos, onde cada estágio executa uma operação nos documentos de entrada [↑](#footnote-ref-1)