UNIVERSIDADE FEDERAL DE SÃO CARLOS - *CAMPUS* SOROCABA

CIÊNCIA DA COMPUTAÇÃO

PROCESSAMENTO MASSIVO DE DADOS

PROJETO PRÁTICO

PROFª. DRª. SAHUDY MONTENEGRO GONZÁLEZ

RELAÇÕES DE PLANTAS E ANIMAIS PARA CULTIVO EFICIENTE

GRUPO 12

FELIPE OTTONI PEREIRA

LETÍCIA ALMEIDA PAULINO DE ALENCAR FERREIRA

Fase Intermediária I

Planejamento: Definição da proposta

Sorocaba, 2025

INDÍCE

Sorocaba, 2025

1. DESCRIÇÃO DO TEMA

O plantio companheiro é uma prática agrícola baseada na combinação de diferentes espécies vegetais em um mesmo espaço, visando benefícios mútuos. Essa técnica contribui para a sustentabilidade da produção, pois reduz o uso de insumos químicos, melhora a saúde do solo e promove o controle natural de pragas [1]. As interações entre as plantas podem ser benéficas, quando uma espécie auxilia no desenvolvimento da outra, seja repelindo pragas, atraindo polinizadores ou melhorando a disponibilidade de nutrientes, ou antagônicas, quando uma prejudica o desenvolvimento da outra, seja por competição por recursos, alelopatia ou atração de pragas indesejadas.

Além das relações diretas entre plantas, outros elementos entram nesse ecossistema agrícola, como insetos benéficos, que polinizam ou predam pragas, pragas específicas, que podem ser repelidas ou atraídas pelas plantas, nutrientes do solo, cuja disponibilidade pode ser alterada por algumas espécies, como as leguminosas, que fixam nitrogênio. Nesse sentido, entender essas relações pode fornecer insights úteis para práticas agrícolas mais produtivas, resilientes e sustentáveis, promovendo um manejo mais inteligente do solo e das espécies cultivadas. Dado que essas relações formam uma rede complexa de interações biológicas, o uso de tecnologias de dados é interessante para mapear, analisar e extrair conhecimento desse sistema.

2. OBJETIVO

Nosso projeto tem como objetivo desenvolver um sistema em duas frentes: uma que explore as relações ecológicas entre plantas no contexto agrícola, considerando aspectos como relações de plantas companheiras, plantas antagônicas, interações com pragas, relação com insetos benéficos e fornecimento de nutrientes, como plantas que fixam nitrogênio ou umidade (água). Com isso, nossa proposta visa gerar insights úteis para agricultores, pesquisadores e profissionais da agricultura sustentável, permitindo uma melhor tomada de decisão no planejamento de cultivos consorciados, proteção do solo e aumento da produtividade de forma ecológica. Enquanto a outra frente visa uma análise de mercado agrícola por região, contendo dados e métricas de produção das culturas em anos anteriores, para assim auxiliar a tomada de decisão quanto a qual cultura investir, avaliando o crescimento em cada setor agrícola. Portanto, uma frente visa auxiliar a tomada de decisão sobre qual cultura investir e plantar, enquanto a outra visa explorar as relações entre culturas para auxiliar e otimizar a produção da mesma.

Do ponto de vista técnico, o projeto tem como objetivo explorar o uso de tecnologias de dados para representar, modelar e analisar redes ecológicas , capturando as múltiplas interações entre plantas, pragas, insetos benéficos e nutrientes. Além disso, busca-se demonstrar a viabilidade e a eficiência da integração entre o banco de dados orientado a grafos Neo4j, o banco de dados orientadora documentos

MongoDB e o ambiente de processamento distribuído Apache Spark, aproveitando os pontos fortes de cada tecnologia na manipulação, análise e extração de conhecimento a partir de grandes volumes de dados inter-relacionados.

3. TECNOLOGIAS

Neo4j

O Neo4j é um banco de dados **NoSQL** orientado a grafos, projetado para armazenar e consultar dados altamente interconectados. Diferente dos bancos relacionais tradicionais, ele organiza as informações em nós (entidades) e arestas (relacionamentos), permitindo uma **visualização clara** e uma representação muito mais natural de redes complexas, como redes sociais, cadeias de suprimentos, mapas de rotas e, no nosso caso, redes ecológicas agrícolas. Em **Modelos Relacionais** modelar essas conexões com tabelas e FKs seria extremamente complexo, com **vários *joins*** recursivas e **tabelas** auxiliares. As relações entre plantas, pragas, insetos benéficos e nutrientes são naturalmente interconectadas, formando uma estrutura de rede que se encaixa muito bem no modelo de grafos, Nesse contexto, o armazenamento em Neo4j atuará na frente que visa explorar as relações entre as plantas para utilizar práticas como plantas companheiras e rotação de cultura para otimizar a produção de alguma planta. Tendo esse objetivo como base, os nós representam entidades como plantas, pragas e nutrientes, enquanto as arestas representam as relações entre eles, como “atrai”, “repele”, “fornece nutriente”, “beneficia” ou “antagônica de”.

Escolhemos utilizar o Neo4j uma vez que ele se destaca por oferecer uma abordagem **altamente intuitiva** e eficiente para modelagem desse tipo de dado relacionado, permitindo que as relações sejam navegadas de forma rápida, mesmo em redes densamente conectadas. Além disso, por meio da **linguagem Cypher**, é possível realizar consultas sofisticadas sobre padrões de conexões, **encontrar cadeias de relações**, detectar comunidades e realizar análises estruturais da rede. **As relações seriam o foco**, sendo propriedades navegáveis diretamente, permitindo **consultas baseadas em caminhos**. Essa capacidade de **compreender** não apenas os dados, mas também **as relações entre eles**, é essencial para extrair insights significativos no contexto do plantio companheiro e do manejo ecológico.

Além disso, a **fonte de dados é semi-estruturada**, com **alta variabilidade e pouca padronização**. Nossos dados encontrados não possuem uma padronização exata, tendo uma **flexibilidade no esquema** - há campos com listas e multivalorados, campos “mistos” (onde os valores neles, uma hora tem um certo nível e outra hora outro), e comentários despadronizados, como mostra a Figura 1. Utilizar um modelo relacional exigiria múltiplas tabelas de normalização e haveria presença de campos nulos e falta de flexibilidade.

VEGETABLE	FRIENDS	FOES
Artichoke	Brassica family	
Asparagus	basil, marigold, nasturtium, parsley, tomato	allium family
Bean-Bush	beets, celery, corn, cucumber, nasturtium, peas, radish, strawberry, summer savory	allium family, fennel
Bean-Pole	carrot, catnip, celery, chamomile, corn, cucumber, garlic, marigold, nasturtium, oregano, peas, potato, radish, rosemary, spinach, squash, summer savory	allium family, beets, brassica family, fennel
Beets	Allium family, beans (bush), lettuce, brassica family, tomato	beans (pole)
Broccoli	artichoke, beets, candy tuft, catnip, celery, chamomile, cucumber, dill, garlic, hyssop, mint, nasturtium, onion, oregano, pennyroyal, peppermint, potato, radish, rosemary, sage, southernwood, thyme, wormwood,	basil, beans (pole), peas (snap), strawberry, tomato

Figura 1: Exemplo dos dados encontrados como fonte.

MongoDB

O MongoDB é um banco de dados NoSQL orientado a documentos, ideal para armazenar grandes volumes de dados com estruturas flexíveis e dinâmicas. Ao contrário dos bancos relacionais, que exigem esquemas rígidos e normalização em diversas tabelas, o MongoDB permite armazenar documentos em formato JSON, o que facilita a modelagem de dados com estruturas aninhadas, listas e campos multivalorados. No contexto do nosso projeto, essa flexibilidade é essencial para representar a produção agrícola de diferentes culturas em diversos países e anos, já que cada documento pode conter informações variadas como culturas, métricas de rendimento e produção, etc. Ou até mesmo pode conter esquemas inconsistentes, campos ausentes vindo da fonte de dados. Portanto como o MongoDB permite campos heterogêneos e esquemas flexíveis e dinâmicos, ele acaba sendo uma ótima opção.

Essa abordagem documental é particularmente vantajosa para a frente do projeto voltada à análise de mercado e produção agrícola por região. Cada documento pode representar os dados de produção de um país em um determinado ano, e dentro dele é possível armazenar uma lista de todas as culturas cultivadas nesse período, incluindo suas respectivas métricas de produção, rendimento por hectare, área plantada e outras condições do plantio. Com isso, o modelo favorece consultas agregadas e comparações entre culturas e regiões sem a complexidade de múltiplos joins ou tabelas auxiliares, como seria necessário em modelos relacionais tradicionais. Além disso, a estrutura flexível permite evoluir o modelo com o tempo, incluindo novos campos conforme novas variáveis se tornam relevantes.

Nesse contexto, o uso desse modelo documental fortalece a capacidade de análise da produção agrícola por região, permitindo gerar insights e consultas que mostram, por exemplo, a evolução do

rendimento de uma cultura ao longo dos anos, o comparativo entre diferentes estados ou países em determinado período, etc.

Apache Spark

O Apache Spark é uma plataforma de processamento de dados distribuído, projetada para lidar com grandes volumes de dados de forma eficiente, rápida e escalável. Seu modelo é baseado em processamento em memória, o que permite realizar operações analíticas e transformações de dados com alta performance.

No nosso contexto o Spark será utilizado para ETL: Extração de dados de múltiplas fontes, Transformação dos dados (limpeza, padronização, enriquecimento e organização) e Carga (Load) dos dados processados para o Neo4j, onde serão representados no formato de grafo, com nós e relações. Mas também é interessante para análises em larga escala, como realização de agregações complexas ou cálculo de estatísticas, como o número de plantas que repelem determinada praga, junto ao Neo4j com a linguagem Cypher. Além disso, a existência de conectores nativos entre Spark e Neo4j permite uma integração direta e eficiente, viabilizando um fluxo de dados dinâmico entre os dois ambientes.

Fontes de dados

Nossas fontes de dados, que encontramos até o momento, são tabelas [2] e artigos científicos sobre agroecologia, websites especializados em plantio consorciado [3][4] e datasets complementares sobre pragas e insetos benéficos, que ajudam a mapear quais espécies são repelidas ou atraídas.

4. MODELAGEM

Fluxograma

Nesta seção temos o fluxograma do projeto, figura 2, com as tecnologias e em que etapa serão utilizadas. O processo inicia com a coleta de fontes de dados diversas, que são então submetidas a um processo de ETL, com o Spark, para limpeza e organização, com os dados sendo inseridos no Neo4j para serem armazenados. A análise dessas relações é realizada por meio de consultas com a linguagem Cypher, permitindo a extração de padrões e insights a fim de auxiliar agricultores e pesquisadores na tomada de decisões para práticas agrícolas mais sustentáveis e produtivas.

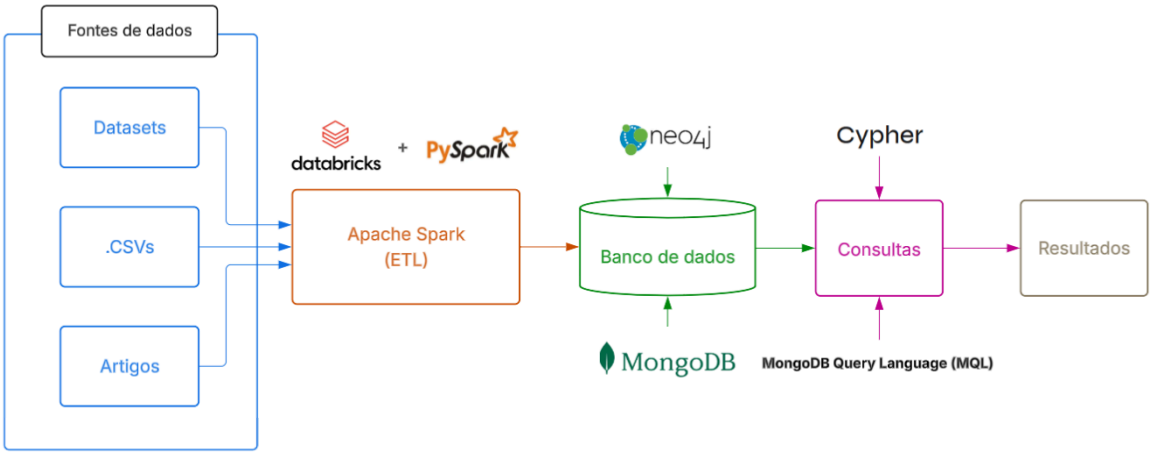

Figura 2: Fluxo de dados e processamento do projeto.

Esquema no Neo4j

Nosso esquema, figura 3, é um modelo orientado a grafo, desse modo, apresenta cinco tipos principais de nós (entidades) e as arestas (relações) que os conectam.

Nós:

- `Planta:` diferentes espécies de plantas envolvidas (ex: Brócolis).
- `Gênero:` gênero botânico ao qual as plantas pertencem (ex: Allium).
- `Categoria:` classificação mais ampla que pode conter gêneros ou plantas (ex: Frutas).
- `Animal:` animais, que podem ser pragas ou polinizadores, por exemplo.
- `Mecanismo:` efeito ou processo resultante de uma interação (ex: fixação de nitrogênio, aumento da umidade).

Arestas (Relações):

- `Planta` ↔ `Planta`: uma planta `"AJUDA"` ou `"ATRAPALHA"` outra planta, indicando relações de companheirismo ou antagonismo.
- `Planta` ↔ `Animal:` uma planta `"ATRAI"` ou `"REPELE"` certos animais.
- `Gênero` ↔ `Gênero:` um gênero `"AJUDA"` ou `"ATRAPALHA"` outros gêneros.
- `Gênero` ↔ `Planta:` um gênero `"AJUDA"` ou `"ATRAPALHA"` uma planta, ou o contrário.
- `Gênero` ↔ `Animal:` um gênero `"ATRAI"` ou `"REPELE"` animais.
- `Gênero` → `Mecanismo:` um gênero `"OFERECE"` um determinado mecanismo.
- `Planta` → `Mecanismo:` uma planta `"OFERECE"` mecanismos.
- `Categoria` → `Gênero/Planta:` uma categoria `"CONTÉM"` gêneros ou plantas.

Com este modelo queremos visualizar as interconexões e dependências entre diferentes elementos do ecossistema agrícola, facilitando a análise de como eles interagem e influenciam uns aos outros.

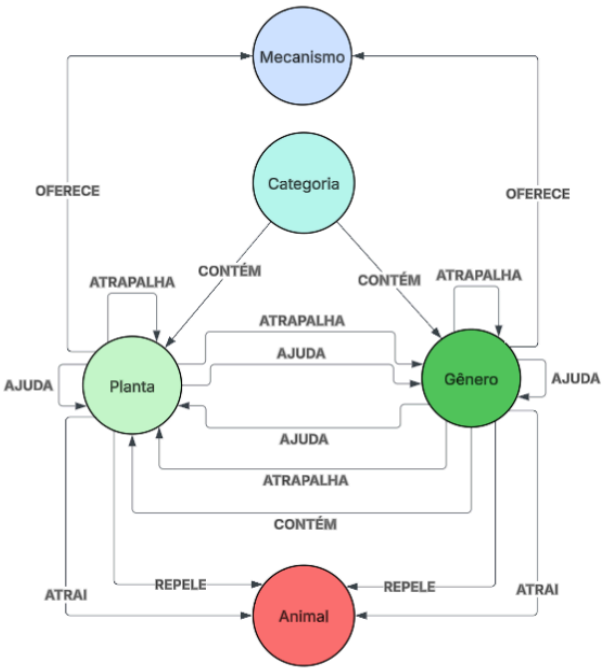

Figura 3: Esquema do modelo de dados.

Esquema no MongoDB

Cada documento representa a produção agrícola de um país em um determinado ano, reunindo em um único registro todos os dados relevantes referentes àquele período e localidade.

Dentro de cada documento, é armazenado o nome da região geográfica, o nome do país, o ano da produção e uma lista de culturas que foram cultivadas nesse país naquele ano. Essa lista de culturas é representada por um array de objetos, onde cada objeto descreve uma cultura específica com informações detalhadas de sua produção.

5. CONSULTAS (IDEIAS INICIAIS)

- ➔ Quais plantas ajudam outras plantas que repelem a praga X?
- ➔ Quais plantas da mesma categoria se atrapalham?
- ➔ Quais as plantas que mais ajudam as plantas da categoria Y? (ranking: top 3)
- ➔ Quais os gêneros de planta que oferecem maior variedade de mecanismos (analisar relação de mecanismo com gênero e com planta daquele gênero)

- ➔ Para cada planta, liste quantas ou quais plantas ela pode alcançar ajudar com até 2 saltos
- ➔ Qual o menor caminho entre a planta A e a planta B, usando relação AJUDA entre plantas (determinar a ordem em que elas devem estar dispostas umas com as outras)
- ➔ Plantas que oferecendo o mecanismo P ajudam outras plantas?
  (Pode identificar tanto plantas companheiras quanto rotação de cultura)
- ➔ Quais plantas atrapalham outras plantas atraindo a praga Z?
- ➔ Quais plantas ajudam outras plantas atraindo um animal (benéfico)?
- ➔ Qual a produção total da cultura X no Brasil no ano Y?
- ➔ Calcular a produção total de uma categoria no Japão?
- ➔ Listar o top 5 de culturas cultivadas nos países da América do Sul nos últimos 3 anos.

6. FONTES

1. Companion Planting | Portland Nursery. Disponível em: <https://www.portlandnursery.com/veggies/companion-planting>.
2. Rotação de culturas: objetivos, vantagens e desvantagens. Disponível em: <https://brasilescola.uol.com.br/geografia/rotacao-culturas.htm>.
3. WIKIPEDIA CONTRIBUTORS. List of companion plants. <https://en.wikipedia.org/wiki/List_of_companion_plants#>
4. HTTPS://WWW.FACEBOOK.COM/MARTHASTEWART. Companion Planting Is the Key to a Thriving Vegetable Garden—Here's How to Pair Varieties to Deter Pests and Attract Pollinators. Disponível em: <https://www.marthastewart.com/8379510/companion-planting-guide>.
5. 14 Vegetables You Should Never Plant Together. Disponível em: <https://www.marthastewart.com/vegetables-to-never-plant-together-8425391>.
6. GOVERNMENT, N. T. Companion planting. Disponível em: <https://nt.gov.au/environment/home-gardens/companion-planting>.

Sorocaba, 2025

7. MOMENI, M. Crop Production. Disponível em: <https://www.kaggle.com/datasets/imtkaggleteam/crop-production>. Acesso em: 18 jun. 2025.

8. HUANG, S. Maintain a Companion Plant Knowledge Graph in Google Sheets and Neo4j | Towards Data Science. Disponível em: <https://towardsdatascience.com/maintain-a-companion-plant-knowledge-graph-in-google-sheets-and-neo4j-4142c0a5065b/>. Acesso em: 18 jun. 2025.

9. COMPANION_PLANT_WIKIPEDIA. companion_plant_wikipedia. Disponível em: <https://docs.google.com/spreadsheets/d/1U4K93EeOU6V4SZ9AgI3wOeV4kgdW9TmKSY_pIypDA-A/edit#gid=0>. Acesso em: 18 jun. 2025.

10. PATEL, Y. Vegetables Cultivation Data Exclusive. Disponível em: <https://www.kaggle.com/datasets/ysthehurricane/vegetables-cultivation-data-exclusive?select=vegetablecropNPK.csv>. Acesso em: 18 jun. 2025.