Ciência de Dados

Fluxograma

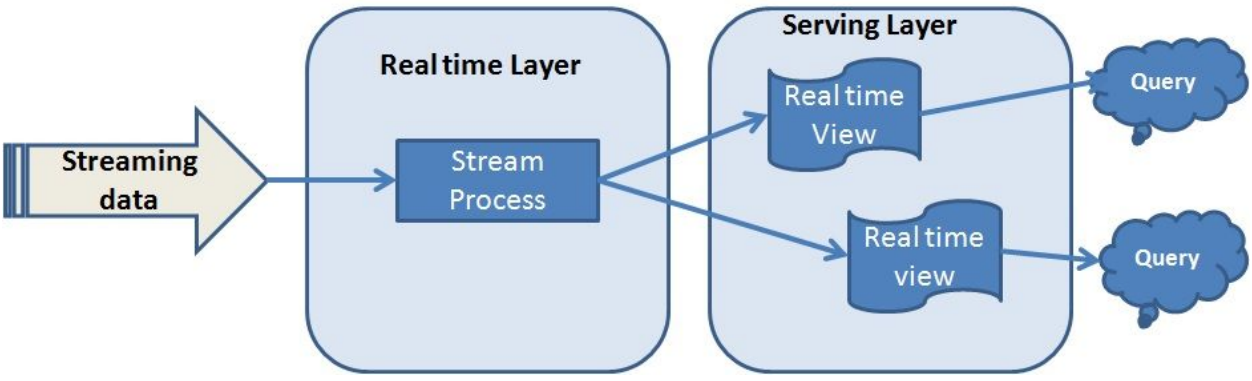
Trabalho 2

Equipe - Abner Lima - 398067, Luan Carvalho - 400583, Marcos Vinicius - 400685

INTRODUÇÃO

O advento do Big Data e suas ferramentas, uma das correntes que teve maior ascensão, foi a da ciência de dados. Dentro do seu ecossistema, duas arquiteturas foram desenvolvidas Lambda e Kappa. Dentre as duas abordaremos somente a primeira, devido a, atualmente, ela estar sendo mais utilizada no mercado e por conseguinte mostraremos como se dá o trabalho do cientista de dados dentro dessa arquitetura.

ARQUITETURA KAPPA



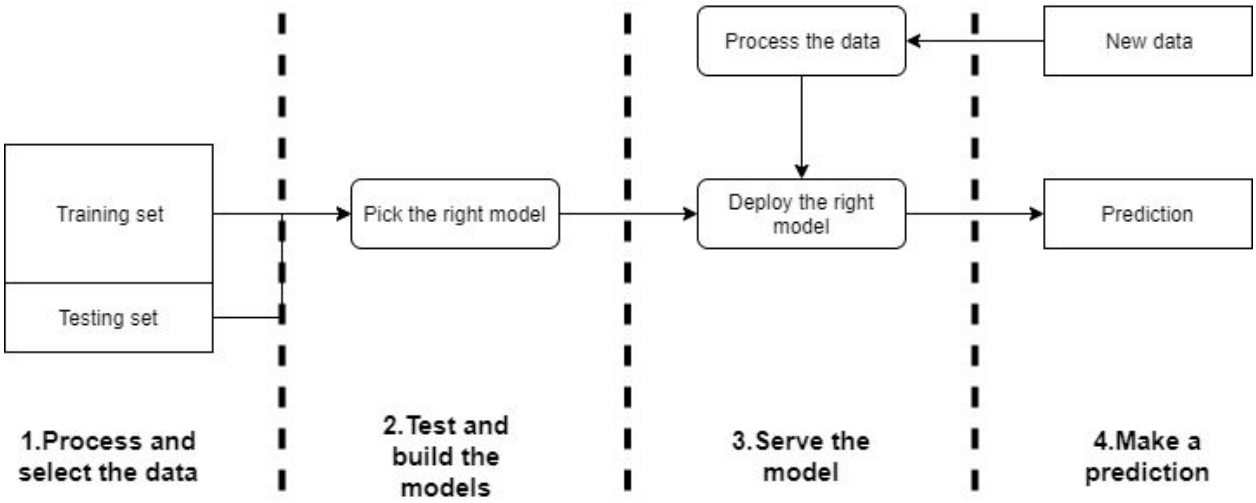
1. Streaming Data: Representa a ingestão dos dados no fluxo, dados esses que podem ser provenientes de diversas fontes. Geralmente para armazenar tudo isso sao utilizadas tecnologias de mensageria, como, o Apache Kafka.
2. Real Time Layer: Essa camada do diagrama, representa o primeiro processamento desses dados dentro do fluxo. Ou seja, é realizada uma primeira limpeza dos dados, junção de dados de múltiplas fontes, envio dos dados para outros microsserviços, entre outros.
3. Serving Layer: Nesta camada, temos um segundo processamento dos dados, os quais a partir daí geralmente servem dados para bancos de dados, data warehouses, entre outros, ou seja, são dados que já estão em um nível de limpeza para serem utilizados por camadas de negócios

DATA SCIENCE NA ARQUITETURA KAPPA

Dentro dessa arquitetura, o cientista de dados pode atuar pegando os dados das duas últimas camadas, real-time e serving, dependendo da necessidade do cientista. O cenário para ele pegar o dado da camada real-time, ocorre quando o cientista quer pegar um dado que ainda não está tão limpo, justamente para analisar as diversas relações desses dados em múltiplas fontes descobrindo certas correlações que quando os dados já passaram por um outro processo de limpeza, as mesmas já não estão mais tão claras.

Já quando os cientistas interagem com a camada serving, eles buscam um dado que já não exigirá um serviço de limpeza tão grande, estando mais pronto para uma exploração e aplicação de modelos de maneira mais rápida e concisa.

Independente dos cenários, o processo do cientista de dados para realizar seu processo, é o mesmo, o qual, será descrito no diagrama abaixo:



1. Process and select data: Nessa etapa, o cientista, independente da camada de onde ele pega o dado, o mesmo, irá realizar um primeiro processo de limpar seu dado, processá-lo e depois de montado o dataset dividir ele em dados de treino e de teste. Dependendo da camada de onde esse dado é pego, ela pode demorar mais ou menos, mas na maioria dos casos é nessa camada onde ele passará mais tempo, visto que além dos processos supracitados, é aqui também onde ele se depara com algumas lógicas de negócio, cabendo a ele explorar ainda mais sobre o business do problema.
2. Test and build the models: Nessa etapa, já munido dos dataset, o cientista irá explorar os possíveis modelos que se aplicam melhor a aquele problema, fazendo todos os processos de validação e afins.
3. Serve the Model: Nessa etapa cabe ao cientista preparar o seu modelo para receber os dados de produção, para no futuro ter a real ciência sobre a aplicabilidade do seu modelo e fazendo os ajustes finais.
4. Make a prediction: Nessa etapa, o cientista de dados recebe os dados novos de produção em seu modelo, gerando valor a companhia com suas predições.

CONCLUSÃO

Por fim, pode se analisar que o trabalho de um cientista de dados não é simples, principalmente quando analisamos o ecossistema de big data, onde o mesmo além de lidar com uma grande volumetria de dados, ele lida também com dados provenientes de múltiplas fontes.

A Kappa Architecture, é uma das tentativas que companhias que lidam com esses problemas, encontram para padronizar sua linha de produção e facilitar o deploy de seu modelo, visto que

ao lidar com grandes quantidades de dados fica fácil de no meio do processo de ter alguma distopia quando não há um processo tão bem estabelecido.

REFERÊNCIAS

1. https://towardsdatascience.com/be-more-efficient-to-produce-ml-models-with-mlflow-c104362f377d
2. https://www.whizlabs.com/blog/real-time-big-data-pipeline/
3. https://docs.microsoft.com/pt-br/azure/machine-learning/concept-ml-pipelines