Projeto Final de Megadados

Por: Antonio Sigrist e Vitória Camilo

Introdução

O projeto que desenvolvemos consiste em realizar um programa que permite analisar o crescimento e transformação do mercado de transportes ao longo dos anos na cidade de Nova Iorque. A partir de uma base de dados que traz diversos dados sobre as corridas de táxis e aplicativos que ocorreram em Nova Iorque entre os anos de 2015 e 2018 (antes de 2015 não haviam dados sobre as corridas de aplicativo, somente as de táxi), decidimos medir alguns pontos que avaliamos como importantes, como o número de corridas realizadas por cada uma dessas modalidades em um mês determinado, quais foram os horários de pico da demanda, comparar o crescimento do número de corridas de táxi e de aplicativos tanto com o mês anterior quanto com o mesmo período do ano passado e quais foram as regiões da cidade que mais iniciaram e terminaram corridas.

Esse projeto foi pensado para ser usado em algumas possíveis aplicações, como para análise interna de aplicativos de transporte para avaliar a evolução do mercado ao longo dos anos, para fundos que desejam estudar antes de realizar algum tipo de investimento ou para uma análise de como alocar e promover incentivos para motoristas priorizarem trabalhar em horários que a oferta seja equivalente à demanda.

Como input, o usuário deve escolher o mês e o ano que deseja analisar. No entanto, a base de dados de alguns períodos não possui algumas informações sobre aplicativos que utilizamos nos cálculos. Nesse caso, as análises desconsideram aplicativos e analisam só os táxis.

Métodos utilizados para o funcionamento do programa

Escolheu-se o Zeppelin como interpretador para a implementação e rodagem da aplicação pois o mesmo já era familiar das atividades da aula e ele foi capaz de realizar todas as tarefas necessárias através das linguagens Python e Spark.

Utilizou-se o sistema de armazenamento em nuvem S3 da AWS para guardar todos os arquivos necessários para o funcionamento da aplicação, e o EMR, serviço de Elastic MapReducing da AWS, que funciona através de clusters, grupos de instâncias que performam tarefas como a análise de grandes massas de dados mais eficientes, pois as distribui entre as máquinas.

Como boa parte das análises realizadas no projeto dependem de dados calculados a partir dos dataframes principais que estão hospedados na S3, não há a necessidade de acessá-los a toda nova iteração, e fazê-lo só deixará a aplicação mais lenta, portanto foram criadas duas bases de dados adicionais (e bem menores). Para fazer isso, todos os cálculos para obtenção dos dados desejados foram feitos previamente em um arquivo separado e guardados em dois arquivos do formato csv. Um deles contém informações sobre as quantidades de viagens realizadas, e o outro, sobre os horários que essas viagens foram feitas, ambas em seus respectivos meses e anos. Estes dois arquivos estão

guardados no bucket que foi criado pelos alunos para a matéria e são eles que são acessados pelo arquivo principal.

Sobre o dataset

Fornecido pela prefeitura da cidade de Nova Iorque através do site http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, as informações inclusas no dataset que são utlizadas incluem: horário de partida das viagens, id da região em que o táxi ou aplicativo foi pedido, id da região que o passageiro de táxi ou aplicativo colocou como destino.

Sobre o código

Na primeira etapa são escolhidos o mês e ano de desejo do usuário através de dois menus dropdown, e após essa escolha são exibidas, se disponíveis, as seguintes informações: a quantidade de viagens realizadas naquele período por táxis e por aplicativos. O aumento ou diminuição percentual que as viagens de táxi sofreram em relação ao mês anterior e ao mesmo mês do ano anterior. E o aumento ou diminuição percentual que as viagens de táxi sofreram em relação ao mês anterior e ao mesmo mês do ano anterior. Para isso, acessa-se a base de dados criada anteriormente que já contém todos estes dados previamente calculados através da função count() e de fórmulas matemáticas para calcular as taxas. Para acessar cada variável separadamente, referente ao período correto e exibir com formatação adequada, utilizamos queries em SQL e a função collect() do Spark.

Em seguida, ao rodar as próximas células exibe-se um gráfico em que o eixo X contém as horas do dia, de 0 até 23, e o eixo Y contém a quantidade total de viagens realizadas por ambos táxi e aplicativo dentro do período de cada horário durante todo o mês. Com essa informação podemos observar os principais horários de pico das viagens. Os dados são também acessados diretamente do segundo dataframe dos alunos, tendo sido previamente calculados a partir da soma das quantidades de viagens que começaram em cada hora, e gravados lá. Para a exibir o gráfico utilizou-se a biblioteca do python matplotlib, e para a formação dos dados plotados, foi preciso criar na mão uma lista com os horários, e gravar no dataset as contagens das viagens por ordem de hora.

Por fim, exibimos quais foram as regiões da cidade em que a maior parte das corridas de táxi começaram e terminaram, e o mesmo para o aplicativos. Para isto, acessa-se uma outra base de dados que contém todas as regiões de Nova Iorque com IDs que também estão presentes na tabela dos dados principais, permitindo relacionar as duas. Conta-se quantas viagens partiram e chegaram em cada região através da função count() e exibe-se o nome da que tiver maior quantidade.

Conclusão

Já era esperado observar um crescimento gradual das viagens de aplicativo com o passar do tempo, mas foi muito interessante poder quantificar este crescimento e o resultado foi impressionante. De janeiro de 2015, o primeiro dataset disponível, até Junho

de 2018, o dataset mais recente, houve um aumento de 768% na quantidade de viagens realizadas através de aplicativos, enquanto para os táxis, houve uma diminuição de 31,65%.

Entre os horários de pico observou-se certa constância. Entre o passar dos anos o horário de maior número de viagens altera-se pouco entre as 18 e as 19 horas. Pela manhã o horário mais procurado é o das 8 horas.

Quanto às regiões mais populares para início e destino das viagens, observa-se que os táxis costumam a sair e chegar em zonas mais nobres da cidade como Midtown e Upper East Side, enquanto os aplicativos são mais comuns em regiões mais populares ou turísticas como East Village e Governor's Island/Liberty Island, ou em regiões desconhecidas pelo dataset (residenciais, por exemplo), representadas por NA.