Então, o que vamos fazer a seguir

é que vamos ver como construir

um pipeline de processamento de fluxo para dados de tráfego em tempo real.

Então, vamos usar isso como exemplo

mas usaremos isso para ver como o processamento de fluxo em feixe se parece.

Então, o cenário aqui é que em uma estrada em San Diego,

nós temos um monte de sensores,

os sensores são colocados um por pista,

e você tem esses sensores em intervalos periódicos ao longo da rodovia.

Então, esses são seus sensores de tráfego

e eles estão medindo e enviando informações para você

sobre o tráfego que está passando por esse sensor.

Então, para processar esses dados,

vamos criar um pipeline de streaming.

A criação de um pipeline de streaming está apenas criando outro pipeline no Dataflow.

Analisamos isso no curso de análise de dados sem servidor.

Então, onde criamos um pipeline de fluxo de dados.

Então, vamos em frente e criar o pipeline,

e dizemos pipeline.create.

Mas esta opção, a única mudança é que você define um streaming para ser verdade. É isso aí.

Então, você define o streaming como verdadeiro

e o Dataflow sabe que agora você não está fazendo lote,

você está realmente computando stream,

mas dentro de um pipeline de streaming,

pode haver coisas que são agregações globais,

e esses seriam em lote.

Então, por exemplo, digamos que estamos levando nossos eventos de trânsito,

nós estamos recebendo as mensagens,

extrair as mensagens e gravá-las no BigQuery.

Então, a fonte é PubSub,

a sincronização é BigQuery. Como isso parece?

Bem, essencialmente você está dizendo PubSubIO ler seqüências de caracteres a partir dele,

do tópico, e essa é sua primeira transformação p.

Isso vai te dar um conjunto de cordas,

e você vai pegar essas cordas

e essa string, você basicamente irá converter e extrair a informação,

e faça um objeto de informação de faixa.

Então, neste momento agora,

você obtém uma coleção P de informações de pista.

Agora, você pega essas informações de coleção de pistas,

estou chamando as condições atuais,

e então podemos convertê-los em uma linha de tabela para o BigQuery,

e podemos dizer BigQueryIO.writeTableRows.

Então, essa parte é semelhante a todo pipeline do Beam que nós escrevemos.

É bem idêntico, a única diferença

foi que tínhamos definido streaming para ser verdade em opções,

e, em vez de ler o texto,

normalmente nós estávamos lendo TextIO,

mas agora estamos lendo o PubSubIO.

Então, essa é praticamente a diferença.

Portanto, o Dataflow é uma ótima maneira de trabalhar com o PubSub.

PubSub lembrar é um serviço de entrega global de baixa latência,

é um barramento de mensagens

mas há alguns problemas.

Não garante a ordem das mensagens

e porque só lhe dá pelo menos uma garantia de entrega,

é possível que você tenha uma entrega repetida,

que a mesma mensagem seja entregue algumas vezes.

O fluxo de dados de processamento de fluxo é responsável por isso.

Número um, quando você está computando agregados,

vai funcionar com mensagens fora de ordem,

porque você não está embaralhando no momento em que você a recebeu,

mas no momento em que a mensagem foi publicada,

e por isso funciona automaticamente com o valor de nossas mensagens.

Lembre-se também que sempre que publicamos no PubSub,

nós tivemos um ID de mensagem.

O Dataflow sabe sobre esse ID de mensagem,

por isso, rastreia qual ID de mensagem já foi processada.

Então, se PubSub acontece para reenviar uma mensagem, não há problema,

O Dataflow vai essencialmente usar esse ID interno do PubSub para desduplicá-lo,

remova-o e acabe processando as mensagens apenas uma vez.

Então, o Dataflow sabe sobre o PubSub,

e porque sabe sobre PubSub,

Juntamente com PubSub pode dar-lhe exatamente uma vez processamento, com garantias.

Assim, você pode impor os únicos que lidam com o fluxo de dados automaticamente.

É só assim que funciona.

Mas isso só funciona com mensagens que entram no PubSub.

O que acontece se a pessoa publicar no PubSub?

A pessoa que publica no PubSub poderia publicar várias vezes.

Então, isso pode acontecer por exemplo,

se você está recebendo dados de um sensor,

o sensor não sabe nada sobre PubSub,

então é só mandar seus pedidos,

e às vezes o sensor desce

volta a subir

volta a subir

e não se lembra se enviou uma mensagem particular para você ou não.

Por isso, reenvia essas mensagens.

Agora, no que diz respeito ao PubSub,

recebeu uma nova mensagem

então vai dar um novo ID de mensagem.

Então, isso não vai funcionar em termos de desduplicação,

porque PubSub acha que estas são duas mensagens independentes.

Então, a linha de fundo é,

se a pessoa que publica no PubSub puder publicar várias vezes,

então quando você publica no PubSub,

adicione um rótulo exclusivo às suas mensagens.

Então, lembre-se que quando publicamos no PubSub,

podemos especificar atributos.

Então, especifique um atributo,

e neste caso,

meu atributo vai ser chamado meu ID,

obviamente pode ser chamado de qualquer coisa,

e quando você está lendo PubSub no Dataflow,

dizemos ao Dataflow que existe um rótulo de identificação

e esse é o rótulo de identificação.

Então, agora, o Dataflow, em vez de digitar o ID interno do PubSub,

irá agora desligar este atributo em vez disso e fazer

Certifique-se de que qualquer ID específico seja processado apenas uma vez.

6:31

Então, de volta, agora sabemos como pegar nossos dados e apenas apará-los.

Ok, isso é processamento de elemento por elemento.

Assim, dissemos que não é tão interessante assim

E se precisarmos fazer agregações?

Vamos dizer, queremos calcular a velocidade média.

Se você quiser calcular a velocidade média em um fluxo de dados ilimitado,

Isso não faz sentido.

Qual é a média sobre algo que não tem começo nem fim?

Você precisa pensar em termos de uma média de um limite dentro de uma janela de tempo.

Você tem que dizer, a média de todo o tráfego

que aconteceu na estrada nos últimos cinco minutos,

ou entre as 7 e as 7:05.

Então, você precisa fornecer essa janela de tempo,

e a janela de tempo que você está fornecendo é uma hora do evento.

A hora em que isso aconteceu.

Então, neste caso,

estamos basicamente dizendo que as condições atuais se aplicam,

e o que estamos fazendo?

Estamos basicamente tomando as nossas condições atuais,

que é o nosso elemento por fluxo de elemento e aplicando uma janela de tempo a ele.

Esta janela de tempo que estamos aplicando é uma janela deslizante,

está desligado cinco minutos a cada 60 segundos.

Então, estamos deslizando uma janela de cinco minutos ao longo da cadeia de tempo.

Por si só, apenas especificando uma janela no evento,

não faz nada para isso.

É só que você está especificando aqui uma janela.

A principal coisa acontece quando você faz uma agregação.

Então, depois de aplicar a janela,

em algum lugar mais tarde no pipeline,

você vai fazer um agregado.

Então, nesse caso, o agregado que estamos fazendo,

é uma média por tecla.

Então, você precisa de algum tipo de chave groupby,

e agora isso significa que você está fazendo

vai ser um meio de apenas os eventos que aconteceram naquela janela de cinco minutos.

Então, essa é a coisa chave.

Então, você aplica uma janela,

e então você aplica o agregado.

Então, para calcular a velocidade média nos dados de streaming,

precisamos vincular esse cálculo dentro de uma janela de tempo.

Então, você é obrigado a isso dentro de uma janela,

e então você calcula o que quer que você queira calcular.

Nesse caso, queremos calcular uma média por sensor,

e porque queremos calcular uma média por sensor,

nós agrupamos por sensor.

Então, basicamente dizemos

me pegue a chave do sensor,

me pegar a velocidade, saída de um par de chaves da chave e da velocidade,

e fazemos uma média por tecla.

Então, vamos obter a média por sensor porque esta chave neste caso é um sensor.

Mas espere um segundo, o que acontece com os dados atrasados?

Então, nós temos uma janela que é uma janela de cinco minutos,

está fazendo isso a cada 30 segundos,

ou a cada 60 segundos,

mas o que acontece se alguns dados chegarem três minutos atrasados?

E nós já fechamos a janela

calculamos a média

e nós nos concentramos?

Então, por padrão,

10:03

o gatilho acontece quando você tem o que é chamado de marca d'água,

e vamos falar sobre o que é uma marca d'água.

Então, quando isso acontecer uma vez na marca d'água,

e eu vou falar sobre o que é uma marca d'água depois,

e a média é escrita.

Mas observe a marca d'água que você tem

você pode ter alguns dados atrasados ​​depois dessa marca d'água,

e se você receber dados atrasados,

essa média será recomputada.

Então, toda vez que você consegue um novo álbum,

nós vamos recalcular este meio,

e nós vamos fazer o stream novamente.

Então, o Dataflow vai lidar com isso para nós.

Para a configuração de disparo padrão,

A maneira como você se lembra disso é que o gatilho padrão é um erro do lado da precisão.

Queremos ter certeza de que seu resultado está correto.

Então fechamos a janela

nós escrevemos a média,

nós dizemos que aqui está uma média

isso é um resultado especulativo,

vem uma coisa nova que precisa fazer parte disso,

vamos recalcular e vamos escrevê-lo novamente.

Obviamente, você quer ser capaz de mudar isso, e você pode.

Mas vamos falar um pouco sobre como você muda isso.

Então a maneira de mudar isso, seja de

faça da maneira mais precisa para fazer algumas compensações,

é preciso entender a ideia do que é uma marca d'água,

e o que é desencadear.