muito bem muito bem vamos lá para mais um falando e andando e hoje mais uma vez aqui ó diretamente das ruas de amisterdan aqui na Holanda durante o inverno europeu céu nublado mas nós temos que caminhar e seguir em frente certo Afinal nós temos que falar sobre um assunto super importante um assunto que impacta diretamente o bolso aí de quem usa kubernetes na nuvem certo hoje é dia de falar sobre desperdícios de recursos assunto super importante enquanto nós vamos passando aqui ó na frente de um coffee shop que vende aquelas ervas especiais holandesas certo recentemente aí saiu um relatório de benchmark aonde os caras analisaram mais de 2100 empresas que estão utilizando kubernetes na nuvem vou te falar o seguinte hein mano os números são assim assustadores certo nesse vídeo nós vamos fazer aí um Deep Dive no relatório que eu gastei um tempinho lendo o relatório mas caso vocês queiram ver o relatório tá o link aqui embaixo na descrição para que vocês possam ir lá no site fazer o download lá no site da Cash certo você vai lá e faz o download do relatório aí você pode ler por si certo por você mesmo mas olha só a ideia é a gente pegar e vamos entender aí como que nós estamos desperdiçando tanto dinheiro com clusters mal configurados e o mais importante o que podemos fazer para resolver isso então bora lá entender um pouquinho sobre as coisas que eu vi lá naquele relatório antes de mergulhar lá nos dados vamos entender o contexto desse relatório certo a Cash ai analisou mais de 2100 organizações que usam kubernetes que nem eu falei os dados são de Janeiro a Dezembro de 2024 foram analisados lá aws gcp e azure ão as três principais Cloud providers aí que nós temos certo só Entraram na análise clusters que tem pelo menos 50 cpus Então são clusters ali e médio para cima um detalhe importante todos os

dados foram coletados antes dessas empresas implementarem qualquer otimização lá com a galera da cas Então vamos lá falando de uma fotografia real de empresas que estão utilizando o kubernetes ainda lá na sem a otimização do jeito que a vida é certo e Vale destacar que os dados de preço vem diretamente das apis públicas lá de inventário dos provedores que a gente falou né da WS da azure e também da gcp certo ah com amostras a cada 60 segundos aí é os dados lá do relatório ou seja tem dados lá e são bastante precisos beleza bem vamos começar primeiro lá falando com os números que eu achei mais chocantes lá do relatório que é a utilização média de CPU nos clusters E ó é apenas 10% sim a utilização média de CPU nos clusters analisados é de 10% e a mesma coisa acontece com a memória certo mas esses 10% aqui representa ainda uma queda de 3% em relação a 2023 quando a média era de 13% já a utilização de memória que nem eu comentei é de é um pouco melhor mas nada muito além daquilo não é de 23% e esse número aumentou 3% Desde o ano passado Aonde era 20% mas galera ó vamos ser honesto 10% de utilização de CPU é muito baixo significa que 90% do que pagamos lá simplesmente não tá sendo utilizado tá indo pro ralo e vem outro outro dado impressionante
99,94 dos clusters analisados estavam sobre sobre aprovisionador
ah do que necessário do que era necessário ão praticamente todos beleza e não estamos falando ainda de um pequeno assim sobre aprovisionamento não a diferença média dos recursos provisionados é é assim tipo mano é é bizarro era de 40% se eu não me engano pra CPU e
57% pra memória coisa para caramba então basicamente
aí mesmo quando pedimos recursos via sei lá request ou limit lá dentro do

kubernetes ainda estamos pedindo muito muito mais do que estamos utilizando loucura certo ah e olha não importa qual o provedor inclusive certo tanto na WS gcp ou azure apresentaram a mesma utilização assim ou estão bem semelhantes um pouco melhor aqui um pouco pior ali mas na média ali foi basicamente a mesma coisa para que você tenha uma ideia sei lá WS era 10% de CPU gcp era 12% e a azure 88% Beleza então vale a pena a gente tem que pegar e ficar ligado não é o problema né né mais uma vez não é o provedor mas sim o jeito e a complexidade de administrar aí o kubernetes mas agora lá vamos lá vamos para as boas notícias né o uso de instâncias Spot que pode reduzir assim drasticamente os custos do cluster para quem não conhece instâncias Sport né são recursos de computação não utilizado que os provedores oferecem com descontos ali significativos bem significativos mas que podemos e e o o problema é que a gente pode perder essas instâncias do nada assim tá ligado quando o o o provedor precisa dela eles vão lá e pegam de volta mas olha só que legal os números lá do relatório são são bem interessantes os clusters com Mix né de instâncias então instâncias on theem e Spot economiza em média 59 por. já clusters rodando apenas eot a economia é de
77% loucura certo estamos falando aí de reduzir os custos sei lá de duas vezes duas vezes e me certo quando a gente tá utilizando somente Spot loucura muita grana muita grana mas ó se liga Eu falei que as a as instâncias elas podem ser interrompidas a qualquer momento não é isso o relatório traz os dados concretos aí sobre isso olha só a WS ela é a mais volátil 51% das interrupções aconteceram lá na primeira hora e os nodes viv em média sei lá 7 eu acho que era 7.6 horas 7 horas aí em média Beleza o gcp ele tem um um um números intermediários 32% das

interrupções são na primeira hora e a vida ali média vai ser de 13 horas 13 horas 13.8 se eu não estou enganado a azure é a mais estável de todas tá ligado apenas 18% das interrupções foram na primeira hora e a média ali de vida foram mano loucura quase 70 horas 69 horas tá ligado então quase 70 horas de vida média numa no node Spot na azure é algo que assim mano é é algo de Mano de Levar bastante em consideração aí no dia a dia Beleza ainda tem lá falando sobre GPU né falando que a azure lidera a disponibilidade de GPU ah das gpus modernas né que nem as a 100 da vida o gcp oferece boa coberturas da v100 manja e por aí vai então eu não vou entrar em tantos detalhes da GPU porque ainda não é assim o Nossa o nosso dia a dia mas você vê lá a economia né com as distâncias Sport para GPU também da mesma forma é tipo é por volta a eor se não me engano é
90% de diferença no preço daí aws e o gcp vai ficar ali na casa dos 66 67 % loucura loucura loucura Beleza então olha só tem muita coisa legal nesse relatório vale a pena a Cash ai mandou muito bem nesse relatório coletando muita informação mais de 2000 empresas é coisa para caramba certo mais uma vez Lembrando que o link para que você possa fazer o download desse reporte aí né Desse relatório desse bmk tá aqui na descrição e no primeiro comentário fixado vai lá faz o download para que você possa ver com detalhes e ó e fica aí né a coisa que nós temos que ter na mente sempre sobre sobre aprovisionamento aí de recursos né a gente tem que pensar aí sobre a o que nós temos lá e a carga real que nós precisamos tudo isso daí é importante ambientes não produtivos ali também tá ligado rodando à noite Parece coisa boba mas eu conversando aí com o pessoal Sei lá eu acho que era da Globo os caras falaram mano desligando é o ambiente de

desenvolvimento à noite tipo redução de milhões tá ligado um bagulho assim então é a absurdo a forma como nós estamos utilizando os recursos de forma irresponsável tá ligado então Ó para você que gostou do vídeo aí faz lá embaixo lá tudo comenta lá faz os sinaizinhos para eu saber que você fez até agora assistiu até agora o vídeo e no mais é isso vejo vocês nas esquinas da internet é nós curte o vídeo compartilha e vai