Låt flashdisken indexera dina ostrukturerade data

Utnyttjar perioder när processorn är sysslolös till att gå igenom innehållet på disken och generera metadata.

Av Noam Mizrahi, Marvell

Noam Mizrahi är så kallad Marvell Fellow med ansvar för teknik & arkitektur. Hans fokus ligger på beräkningskraft i noderna, distribuerad lagring i näten och AI. Tidigare har han arbetat med att ta fram företagets arkitektur för att flytta stora datamängder mellan servrar och flashdiskar kallad NVMe-over-Fabrics.

Övervakningskameror, sensorer, drönare, uppkopplade bilar, IoT-noder och industriella system – alla genererar de data i en rasande takt. Men förväxla inte data med information. Det är en avgrundsdjup skillnad.

Bara en bråkdel av all insamlad data är värdefull nog att betraktas som en tillgång. Ta till exempel en videokamera. Det kan hända att en inspelning på flera timmar bara innehåller en enda intressant minut och att det under resten av tiden inte händer något som är relevant.

Det är som att gräva efter guld där guldklimparna är information och sanden är data. Förmågan att omvandla data till värdefull information brukar gå under beteckningen analys.

Figur 1 baseras på data från Statista och visar på den fenomenala utbyggnad av datalagring som skett under det senaste decenniet. Prognosen för nästa år pekar på ett behov av åtminstone 42 000 exabyte. Större delen av data som lagras, det finns siffror som pekar på åtminstone 80 procent, är totalt ostrukturerat och därmed svårt att analysera.

Det finns uppskattningar som säger att endast fem procent av allt lagrat data används i analyser. Om det fanns ett sätt att presentera denna ostrukturerade data med hjälp av metadata som förklarade vad det var, skulle mycket mer av det kunna analyseras. Därmed skulle företag och organisationer kunna öka värdet på de data som de sitter på.

ARTIFICIELL INTELLIGENS är en teknik som påverkar alla delar av samhället. Det används för rekommendationer vid e-handel, för översättning mellan språk och av finans- och säkerhetsbransch. Det används även för identifiering av objekt – på sjukvårdsområdet kan exempelvis livshotande cancerceller snabbt upptäckas och identifieras. Trots att områdena är så disparata finns en gemensam nämnare: teknik som kan gå igenom enorma mängder ostrukturerad data (video, text, röst, bilder, med mera) och bearbeta dem så de visar sitt rätta värde.

Vi kan inte bara använda AI för analysdelen utan också för att preparera ostrukturerad rådata genom att tagga det med metadata som på ett enkelt men ändå exakt sätt förklarar vad det är. Databasen kan sedan analyseras av mjukvara i övre lager som skapar användbar information. Alla har väntat på att AI ska hjälpa oss att få ut mycket mer av all den data vi lagrar.

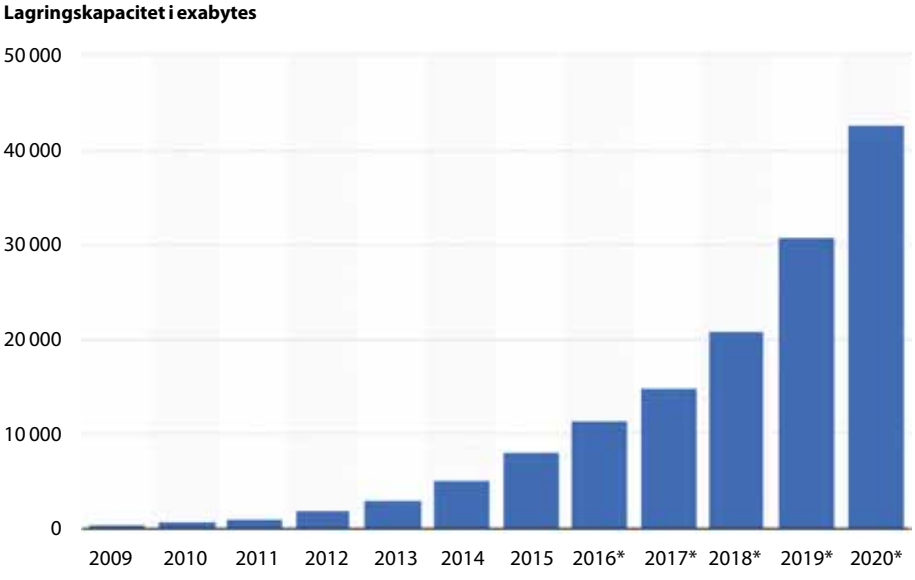
Om vi nu vill generera metadata för att våra analysverktyg ska bli effektivare, och vi har AI som kan skapa metadata till våra enorma mängder ostrukturerade data, är det enda vi behöver göra att flytta all data till platsen där AI-motorn körs. Men är det verkligen rätt väg att gå?

Om vi tittar på var data skapas och var det lagras på fältet och i molnet, blir det snabbt uppenbart att det är väldigt dyrt att flytta data, så det är något som bör undvikas. Även att flytta data inom ett datacenter kostar energi och upptar bandbredd. Tittar man istället på ändnoderna har dessa begränsad energibudget och beräkningskapacitet. Dessutom kan bandbredden för att skicka data till molnet vara begränsad vilket gör det opraktiskt att bearbeta data i molnet.

I BÅDA FALLEN är nyckeln till effektivitet att minimera förflyttningarna av data och istället förlita sig på metadata.

Det skulle bli mycket effektivare om vi kunde ta fram metadata vid källan, exempelvis inuti lagringsenheten. Flashdiskar (SSD) innehåller redan de element som behövs för uppgiften. Processorn används normalt bara för att styra skriv- och läsoperationer men skulle även kunna användas för andra uppgifter, som att märka data. Man kan också komplettera med ytterligare hårdvara, firmware och mjukvara för att lösa uppgiften.

Ett arbetssätt vore att utnyttja perioder när disken är i vila för att skapa metadata. Ett annat att utföra uppgiften i samband med att data skrivs till disken.



Figur 1. Förväntat behov av datalagring från 2009 till 2020.

Fördelarna i bägge fallen är att man sparar energi och pengar samtidigt som fördröjningar och förflyttning av data minimeras och belastningen i nätet sjunker.

Skalbarheten i konceptet gör det möjligt för företag och molnleverantörer att bredda sina erbjudanden med hjälp av artificiell intelligens.

På konferensen Flash Memory Summit i Santa Clara i augusti förra året presenterade Marvell en helt ny AI-processor för flashdiskar som visade hur data kan märkas på ett effektivt sätt utan att man belastar CPU-resurserna och därmed slipper de tillhörande kostnaderna och fördröjningarna. Deltagarna fick se ett vanligt datacenter från Marvell och en processor för SSD-diskar som använde Nvidias öppen källkodslicensierade Deep Learning Accelerator. Den använder en tränad AI-modell kompilerad till ett IP-block som utför inferenser på en stor databas av ostrukturerad data (exempelvis ett videobibliotek) som finns lagrat lokalt på disken. Det adderar metadata till databasen som på ett smidigt sätt representerar data och kan nyttjas för sökningar i databasen.

Om målet är att upptäcka och känna igen objekt eller scener kan AI-motorns inferenser söka igenom videomaterialet på disken och skapa metadata som listar tidpunkterna för när de syns på videon. Tack vare den nya AI-tekniken kan databasen med metadata lagras lokalt på disken samtidigt som den är tillgänglig för extern programvara som vill utföra analyser.

TA SOM EXEMPEL en polismyndighet som letar efter någonting – en person, ett objekt, en händelse – i video som omfattar många timmar. Myndigheten kan ladda en modell tränad för detta mönster och köra inferenser på videon i bakgrunden parallellt på alla diskar medan videofilmen lagras. Allt som påträffas märks upp, vilket gör det enkelt att plocka fram och studera senare.

På motsvarande sätt vore arkitekturen effektiv för att förbättra en chattrobot. Anta att det finns en stor databas med lagrade konversationer som behöver gås igenom för att förbättra chattroboten. Det skulle vara möjligt att mäta om kunder är nöjda eller missnöjda med svaren de får av roboten eller om konversationerna upplevs för korta eller för långa. När man väl tränat en AI-modell som kan identifiera rätt mönster kan man kompilera den till inferensmotorn i disken och låta den gå igenom och märka upp konversationerna.

För tillämpningar som skräddarsydda annonser till beställvideotjänster (VOD), sökningar efter personer eller objekt liksom andra IO-tunga användarfall ger närheten till data en stor prestandaförbättring.

Den teknik med AI i flashdiskens processor som Marvell utvecklat, visar hur nya arkitekturer för datalagring kan implementeras för att hantera olika utmaningar vad gäller bearbetning av data. Genom att ge de flashdiskar som redan finns på marknaden tillgång till extra logik som gör dem betydligt mer intelligenta går det att skapa metadata och taggar som behövs för olika uppgifter direkt vid källan. Det behövs därmed ingen kommunikation till andra resurser dedicerade för uppgiften.

Med AI-acceleratorer integrerade i kostnadseffektiva systemkretsar för flashdiskar blir det möjligt att snabbt utföra analyser. Lösningen kräver dessutom mindre beräkningsresurser och drar mindre energi samtidigt som man inte behöver utveckla nya asicar. Tack vara den programmerbara arkitekturen går det att uppdatera AI-modellerna så att nya användarfall kan adresseras varefter de dyker upp. ■