Publicatie : AG Connect
Datum : 31 okt 2023 Regio : Nederland Pagina : 6 Bereik : 15.600
cm2 : 1350 Frequentie : 8x per jaar AVE : € 12250,00 Oplage : 6.500

TECH & TOEKOMST

SUPER SCIENCE

PETER BONCZ

AMSTERDAM IS WERELDWIJD EEN BELANGRIJKE HUB VOOR INNOVATIE IN DATABASETECHNOLOGIE. ONDERZOEK AAN HET CENTRUM WISKUNDE & INFORMATICA LEVERDE INMIDDELS VIJF TOONAANGEVENDE STARTUPS OP. DE ONTWIKKELDE DATABASECONCEPTEN VORMEN DE BASIS VAN TALLOZE TOEPASSINGEN, VAN BEURSLIEVELING SNOWFLAKE TOT GOOGLE'S BIGQUERY. DATABASES HEBBEN MISSCHIEN EEN SAAI IMAGO, MAAR DAT IS NIET TERECHT.

De ideale database bestaat niet

Amsterdam is de bakermat van database-vernieuwing

HET OP ORDE KRIJGEN VAN DATA IS VANDAAG DE DAG VERUIT DE MEEST TIJDROVENDE EN INGEWIKKELDE KLUS, of het nu gaat om slimme marktanalyses, het maken van diagnoses in de gezondheidszorg, het presenteren van advertenties aan de juiste doelgroep of het trainen van large language models. En dan moeten die data ook nog eens ontsloten worden op een manier dat de applicatie er optimaal gebruik van kan maken. Databases spelen een cruciale rol in dat proces.

De meeste mensen vergeten dat er niet één ideale database is. Integendeel. "Bij het ontwerpen van een nieuwe database-architectuur begin je met het maken van een wensenlijstje. Dan blijkt al snel dat die eigenschappen vaak met elkaar in conflict zijn", zegt Peter Boncz, hoofd van de onderzoeksgroepen Software Engineering, Database Architecture en Information Access van het Centrum Wiskunde & Informatica (CWI) en hoogleraar Large-Scale Analytical Data Management aan de Vrije Universiteit.

Een database kan idealiter veel data bevatten en je moet er snel in kunnen zoeken. Je moet de data continu kunnen updaten. De opslag moet goedkoop zijn en de database moet makkelijk zijn in gebruik. "Het is een gevecht tussen verschillende belangen. Wil je snel kunnen zoeken, ga je de data misschien indexeren. Maar indexen moeten onderhouden worden. Dat vertraagt weer het toevoegen van data. Als het geheel goed schaalbaar moet zijn, moet het op meerdere machines draaien, maar die moeten met elkaar communiceren." Het zijn maar enkele van de conflicten die Boncz opsomt. De oplossingen komen uit de database-architectuur, een vakgebied dat een stormachtige ontwikkeling doormaakt. "Er bestaat een sterke industriële vraag naar de technologie die wij hier in de groepen ontwikkelen. Databasesystemen zijn een miljardenmarkt. Oracle was lange tijd marktleider, maar raakt die positie kwijt aan kleinere bedrijven en een continue veranderend landschap van nieuwe ideeën. Er is veel kennisuitwisseling en samenwerking tussen de industrie en academische onderzoeksgroepen."

VERLEDEN

In Nederland was database-architectuur eind jaren tachtig een specialisme waar in Nederland nog weinig aandacht voor was. De onderzoeksgroep van Martin Kersten, hoogleraar aan de

Beeld: Marike van Pagée

Publicatie : AG Connect
Datum : 31 okt 2023
cm2 : 1350
Regio : Nederland
Frequentie : 8x per jaar
Pagina : 6
AVE : € 12250,00
Bereik : 15.600
Oplage : 6.500

Publicatie : AG Connect
Datum : 31 okt 2023 Regio : Nederland Pagina : 6 Bereik : 15.600
cm2 : 1350 Frequentie : 8x per jaar AVE : € 12250,00 Oplage : 6.500

Universiteit van Amsterdam, was destijds de enige in het veld, dat verder werd gedomineerd door Amerikaanse partijen. Boncz kwam er eigenlijk bij toeval mee in aanraking. “Ik studeerde software engineering aan de Vrije Universiteit en vond operating systems wel sexy, databases niet. Ik kwam in contact met Martin Kersten en raakte door een proefschrift uit zijn groep onder de indruk van de complexiteit van databases.” Zo startte hij zijn eigen promotie-onderzoek waarbij hij gelijk betrokken raakte bij de eerste startup, Data Distilleries. “In die tijd kwam datamining voor het eerst op het netvlies. Vooral verzekeringsmaatschappijen wilden risicomodellen bouwen op basis van echte data in plaats van intuïtie. Martin Kersten zette Data Distilleries op samen met twee oudere PhD-studenten, ik deed later mee omdat ik de databaseman was.” Boncz had tijdens zijn promotie de database MonetDB ontwikkeld, een snel opensource-databasesysteem. “Datamining is een combinatie van statistiek, databases en een beetje AI, in de zin van regels en voorspelende boomstructuren. Ik was toen misschien nog maar twee jaar bezig als PhD-student, maar de early adopters bonkten zo hard op de deur dat ik erin gerold ben.”

Na een paar jaar Data Distilleries lonkte voor Boncz toch de wetenschappelijke uitdaging. MonetDB werd verder ontwikkeld, waaruit de tweede startup Vectorwise ontstond, genoemd naar de gelijknamige database. “We hebben als groep een aantal belangrijke ideeën geïntroduceerd. Bij MonetDB ging het om de column stores, met Vectorwise voegden we daar het vectoriseren van queryverwerking aan toe. Dat wordt toegepast in veel systemen die nu worden gebouwd, van Google’s BigQuery tot Snowflake en natuurlijk ook onze eigen DuckDB. Ook Databricks heeft een systeem gemaakt op basis van vectorized query execution.”

HEDEN

Door de snelle opkomst van AI-toepassingen en meer specifiek machine learning krijgt de groep van Boncz weer andere vragen uit de markt. “Wij hebben altijd dicht op de hardware-ontwikkelingen gezeten en kijken hoe je die vernieuwingen kunt inzetten voor een betere databaseverwerking. De nieuwe hardware wordt meestal niet gemaakt voor databasetoepassingen, maar bijvoorbeeld voor gaming, of om neurale netwerken sneller te kunnen trainen. GPU’s (grafische chips) en TPU’s (Tensor Processing Units, gespecialiseerde chips van Google voor AI-toepassingen, red.) zijn vaak sneller dan een gewone processor.”

> Organisaties verzamelen veel data en als individu moet je dat maar accepteren

DuckDB is de meest recente database-ontwikkeling in de groep van Boncz. Het is een open source in-process databasemanagementsysteem gericht op het verwerken van analytische zoekopdrachten. Rondom deze kern is in Amsterdam de spin-off DuckDB Labs opgezet door Hannes Mühleisen en Mark Raasveldt, twee medewerkers uit de Database Architecture groep van Boncz. Jordan Tigani, de bedenker van Google BigQuery, heeft weer op basis van een samenwerking met Mühleisen de startup MotherDuck opgezet om DuckDB als een cloudservice aan te bieden. Boncz: “Op al deze populaire ontwikkelingen staat een zwaar CWI-stempel. We hebben ideeën aangedragen voor nieuwe manieren om data op te slaan in de cloud en gewerkt aan columnaire compressie die geschikt is voor vectorized query execution.”

Boncz werkt nu tijdens een jaar sabbatical als software engineer bij MotherDuck. “Ik kan weer zelf programmeren en heb veel bijgeleerd over systeemontwikkeling in DuckDB. Bovendien zijn de problemen die MotherDuck probeert op te lossen, inspirerend voor toekomstig onderzoek.”

TOEKOMST

“We zijn nu onder meer bezig met compressiemethoden die beter geschikt zijn voor de nieuwe generaties hardware, dus de GPU’s en de TPU’s. Eigenlijk een klassiek onderwerp, opslag op een laag niveau dataopslag.”

Daarnaast werkt Boncz aan een oplossing voor het privacyprobleem op niveau van dataopslag. “Organisaties verzamelen veel data en als individu moet je dat maar accepteren als je een app wilt gebruiken. Ik wil met gedecentraliseerde data mensen meer controle teruggeven.” Boncz denkt aan methoden waarbij de degene die de gegevens maakt ze onder bepaalde voorwaarden toegankelijk maken. Er zijn meer initiatieven die dat doel nastreven, maar er is toch een duidelijk verschil, benadrukt Boncz. “Een app als IRMA is een standalone app met een bepaalde beperkte functionaliteit. Het leuke van het ontwerpen van een database is dat die generiek moet zijn zodat die voor veel toepassingen gebruikt kan worden. Dat is nog een relatief nieuw terrein.”

Thijs Doorenbosch
is redacteur bij
AG Connect.