

# Arhitecturi Big Data

O arhitectură Big Data este un sistem conceput pentru a gestiona colectarea, stocarea și procesarea unor volume mari de date (denumite în mod obișnuit "Big Data"). Aceste arhitecturi implică adesea utilizarea de instrumente și tehnici specializate concepute pentru a scala și a gestiona complexitățile seturilor de date mari.

Big Data este caracterizat de așa-zisii "3V": volum, viteză și varietate. Volumul de date se referă la cantitatea mare de date generate și colectate, care pot fi în ordinul de petabytes. Viteza se referă la rapiditate cu care datele sunt generate și procesate, de obicei în timp real. Varietatea se referă la diferitele tipuri de date colectate, inclusiv date structurate din baze de date, date nestructurate din rețelele sociale și tot ce se află pe spectrul dintre ele.

Există mai multe componente cheie ale unei arhitecturi Big Data:

Surse de date: acestea sunt sursele din care sunt colectate datele, cum ar fi baze de date, senzori, rețele sociale etc.

Stocarea datelor: Aici sunt stocate datele, de obicei într-un sistem de fișiere distribuit sau într-un "data lake" .

Procesarea datelor: Aceasta este componenta care se ocupă de procesarea și analiza datelor, de obicei folosind instrumente precum Hadoop sau Spark .

Vizualizarea datelor: La nivelul acestei componente are loc prezentarea rezultatelor analizei datelor într-un mod clar și ușor de înțeles, utilizând de obicei instrumente precum dashboard-uri sau grafice.

Securitatea datelor: Responsabilă pentru protejarea datelor și pentru a se asigura că acestea sunt accesate și utilizate într-o manieră sigură și conformă.

Conceptul de Big Data a devenit din ce în ce mai important în ultimii ani, deoarece organizațiile se străduiesc să obțină informații și să ia decizii mai bune pe baza cantităților mari de date generate. Aplicațiile se afla în vaste domenii inclusiv în afaceri, sănătate, guvern și cercetare.

Un model arhitectural Big Data foarte important este MapReduce. Acesta este un model de programare și implementare asociată pentru

procesarea și generarea de seturi mari de date folosind algoritmi paraleli distribuiți într-un cluster. A fost dezvoltat de Google la începutul anilor 2000 pentru a răspunde nevoilor companiilor care se ocupă de cantități mari de date și de atunci a devenit o metodă populară de procesare a datelor mari. Modelul MapReduce se bazează pe ideea de a împărți o problemă în părți mai mici care pot fi procesate simultan în paralel. Acest lucru se face folosind două funcții: o funcție de "map" și o funcție de "reduce".

Un model alternativ celui de MapReduce este high-throughput computing (HTC). Acesta este un tip de calcul care implică procesarea unor cantități foarte mari de date într-un timp de lungă durată, folosind o multitudine de resurse computaționale de mare putere. În principal HTC este predominant preocupat de robustețea și fiabilitatea sarcinilor pe o scară de timp îndelungată. HTC-urile sunt utilizate frecvent în cercetarea științifică, în special în domenii precum genomica, astrofizica și fizica particulelor, care generează cantități uriașe de date. Un exemplu binecunoscut de HTC este modelul computerizat utilizat de Organizația Europeană pentru Cercetare Nucleară (CERN), care operează Large Hadron Collider (LHC). LHC generează cantități mari de date din ciocnirile sale de particule, iar modelul HTC CERN este conceput pentru a procesa aceste date cu validitate mare.

În general, Big Data schimbă modul în care organizațiile funcționează și iau decizii. Are potențialul de a revoluționa industriile și de a rezolva unele dintre cele mai presante probleme actuale. Cu toate acestea, este, de asemenea, important să se ia în considerare mai multe probleme potențiale și critice cu datele mari și tehnologiile actuale pentru colectarea, stocarea și procesarea seturilor mari de date. Unele dintre preocupările cheie includ:

**Confidențialitatea:** Colectarea și analizarea unor cantități mari de date ridică preocupări legate de confidențialitate, mai ales atunci când datele sunt cu caracter personal. Unele companii și organizații folosesc Big Data în moduri care încalcă confidențialitatea oamenilor, cum ar fi colectarea și vânzarea de date cu caracter personal fără consimțământ

**Complexitate:** a lucra cu Big Data implica necesitatea unor sisteme de complexitate ridicată care au nevoie de oameni specifici instruiți să dezvolte și mențină astfel de sisteme

Securitate: Big Data poate fi vulnerabilă la amenințările de securitate, cum ar fi atacurile cibernetice, aceasta poate fi o problemă pentru organizațiile care se ocupă de date sensibile sau confidențiale

Considerații etice: pot exista probleme cu privire la datele utilizate pentru marketing sau publicitate direcționată sau datele pot fi folosite pentru a discrimina anumite grupuri.

În general, în timp ce Big Data are potențialul de a aduce beneficii și perspective enorme, organizațiile trebuie să ia în considerare cu atenție potențialele probleme și să se asigure că folosesc datele mari în mod responsabil.