Big Data – tömeges adatelemzés gyorsan

STADLER GELLÉRT
Oracle Hungary Kft.
gellert.stadler@oracle.com

Kulcsszavak: big data, döntéstámogatás, hadoop, üzleti intelligencia

Az utóbbi években új informatikai „buzzword" jelent meg és tűnt fel: a „big data". A technológia elterjedése és népszerűsége számos okra vezethető vissza. A nagy mennyiségű adatot tároló és feldolgozó rendszerek skálázhatósági problémájára adott sikeres válaszával indult el, de általános alkalmazhatóságának felismerése egybe esett a digitálisan tárolt adatok mennyiségének robbanásszerű növekedésével, és ezzel párhuzamosan ezen adatok elemzési igényének megjelenésével is.

A big data olyan költséghatékony eszközrendszert adott az elemzők kezébe, amely segítségével könnyebben, olcsóbban, gyorsabban lehet új típusú elemzéseket készíteni és ezeken keresztül versenyelőnyt elérni. A rengeteg különböző gyártó és fejlesztő által fejlesztett megoldások felhasználása, integrálása azonban sokszor nem triviális, nehézségekbe ütközik. Ilyenkor segíthet egy olyan szállító, aki a hardver infrastruktúrától kezdve a szoftveren keresztül a teljes megoldás szállítására és támogatására képes.

Az első szakaszban röviden áttekintjük a big data megjelenésének történetét, a következő szakaszban technológia alkalmazásának főbb motivációit vesszük sorba, végül röviden áttekintjük az Oracle – mint az egyik adatfeldeldolgozási szoftverek tekintetében piacvezető gyártó – erre vonatkozó koncepcióját.

1. A big data rövid története

Körülbelül 2012-től kezdve láthattuk egyre többször felbukkanni mind az internetes, mind a hagyományos sajtóban a „big data" fogalmát. Nem kétséges, hogy mára már ez is a mostanában nagyon divatos informatikai „buzzword"-ök egyikévé vált, hasonlóan a korábbi években feltűnt „Web 2.0"-hoz (2006-) és a „Cloud Computing"-hoz (2009-), – lásd az *1. ábrát*.

Pedig a big data története sokkal korábban, már 2003 környékén elkezdődött. Ekkor publikálta a Google az általa használt GFS (Google Distributed File System) leírását [1]. Ez a technológia éppen kapóra jött az Apache Software Foundation keretein belül dolgozó Doug Cutting-nak és csapatának, akik 2002 óta dolgoztak egy Open Source web keresőmotor kidolgozásán (Apache Nutch) [2]. Az akkori technológiai lehetőségeken alapuló rendszerrel 1 millárd oldal adatait tartalmazó kereső rendszer kb. fél millió dolláros kezdeti hardver költséggel és 30 000 dolláros havi üzemeltetési költséggel tudott volna működni. Ezt a magas költséget elsősorban a web kereső és indexelő program által generált óriási méretű fájlok okozták, amelyek kezelése a hagyományos fájlrendszerekben vagy adatbázisokban csak nehézkesen volt megoldható. A nehézkesség és a magas költségek abból adódtak, hogy az akkori technológia skálázhatósága ebben a mérettartományban már nehézkes és drága volt.

A GFS (vagy egy hasonló elven működő más rendszer) pont erre a problémára adott olyan megoldást, amely nagyságrendekkel megkönnyítette az ilyen nagyméretű fájlok kezelését. 2004-ben Cutting és csapata elkezdtek dolgozni egy GFS-hez hasonló fájlrendszer open source keretek közötti megvalósításán, ez lett a Nutch Distributed File System (NDFS). 2004-ben publikálta a Google a MapReduce programozási keretrendszer leírását [3], 2005 elején a Nutch fejlesztők a teljes kereső rendszert átírták MapReduce és NDFS alapokra. Mivel az elkészült rendszer alkalmazhatósága jóval túlmutatott a web keresés problémáján, 2006-ban a technológia továbbfejlesztésére megalapították a web keresés problémájától függetlenített Hadoop alprojektet, ami 2008-ra önálló Apache projekt lett. Ugyanebben az évben jelentette be a Yahoo, hogy a keresőmotorja egy 10 000 processzor magot tartalmazó Hadoop clusteren alapul. Ebben az évben több más cég is bejelentette a technológia éles környezetben történő alkalmazását (Last.fm, Facebook, New York Times).

A big data tehát alapvetően a nagymennyiségű adat kezelésekor fellépő technológiai skálázódási problémára adott válaszként indult. A Hadoop – mint az első, széles körök számára elérhető big data technológia – azért lett olyan sikeres, mert csaknem lineálisan skálázható akár a Petabyte-os nagyságrendig is. Ezután még sokáig ez volt a big data fő alkalmazási területe. Ez önmagában még nem indokolta volna azt, hogy a Web 2.0-hoz vagy a Cloud Computing-hoz hasonló népszerűségre tegyen szert, mivel az ilyen nagy mennyiségű adatot kezelő cégek és alkalmazások száma viszonylag kevés. Valami más is történt, ami ezt a folyamatot igazán elindította a tömegesebb alkalmazás irányába.

A HDFS és a MapReduce technológia alkalmazásának bekerülési költsége alacsony. A Hadoop a párhuzamosított feldolgozás és a magában foglalt adattárolási és feldolgozási redundancia kettősségére épül, nem igényel drága, vállalati szintű szervereket vagy tároló rendszereket, hanem ezt gyakorlatilag PC szintű számítógépek és diszkek alkalmazásával képes lineális mértékben skálázni. Az addig gyakorlatilag egyeduralkódónak számító relációs adatbázis alapú és SAN fájl rendszeres technológia alkalmazásához képest kisebb költséggel tud működni.

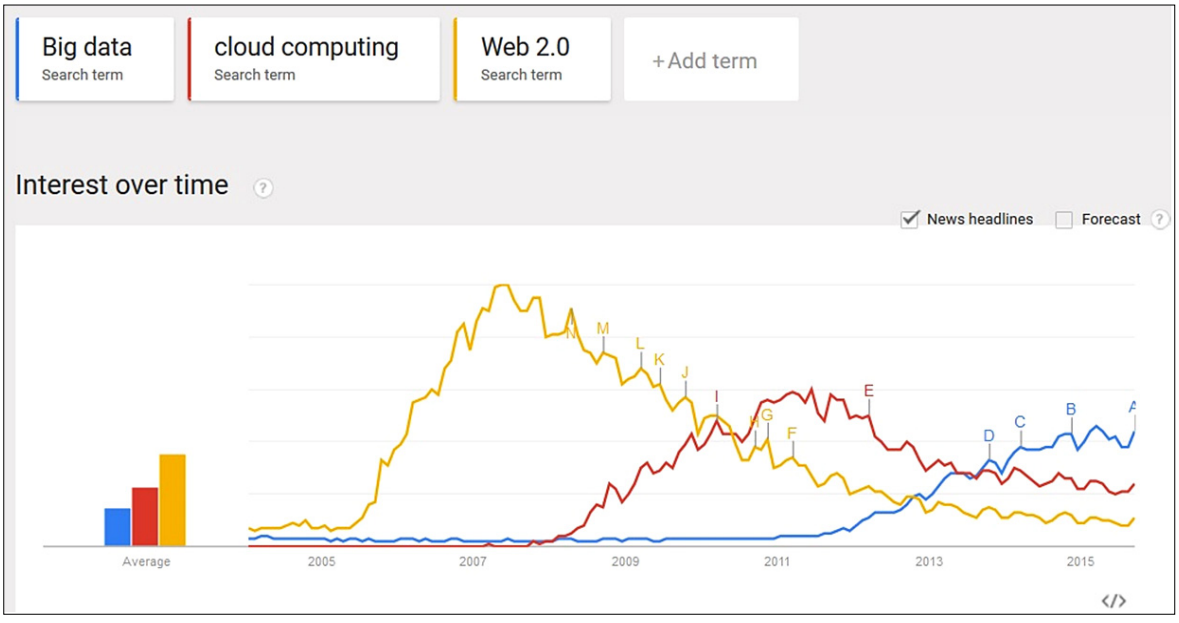
A népszerűséghez viszont más is kellett. Kiderült, hogy nagyon sok olyan elemzési igény van különböző cégeknél és elemzőknél, amihez eddig nem találtak megfelelő kapacitást. Ez részben abból fakadt, hogy a magas tárolási költség miatt nem is tároltak el bizonyos adatokat, más részben abból, hogy bár tárolták az adatokat, de azokat nem tudták kiaknázni, mert a feldolgozáshoz szükséges gépidő/számítási kapacitás nem állt rendelkezésre. Itt nem kell feltétlenül több petabyte-os adatmennyiségekre gondolni. Egy közepes vagy kisebb cégnek már néhány terabyte-nyi adat tárolása is nehézséget okozhat, nem beszélve annak elemzéséről, ami jelentős számítási kapacitást igényelhet. A Hadoop (és így a big data) megjelenésével egyre több olyan adat kezdett el Hadoop cluster-be töltődni, amely korábban elemzés számára elérhetetlen volt.

A szoftver technológia ingyenessége (Open Source) és az igényelt (commodity) hardver viszonylagos alacsony költsége miatt egyre többen kezdtek érdeklődni a Hadoop iránt. A kezdeti gyors elterjedést azonban korlátozta az, hogy a relációs vagy más elterjedt adatbázisokon alkalmazható elemzést támogató technikákhoz képest (SQL, OLAP stb), a hadoop csak egy nagyon alacsony szintű eszközrendszert adott az elemzők kezébe, amelyben minden lekérdezés csak programozással volt megvalósítható. Ez behatárolta az alkalmazók körét, ugyanakkor többen elkezdtek foglalkozni azzal, hogy új, fejlettebb elemzéseket lehetővé tévő technológiákat hozzanak létre. Sok cég kezdett foglalkozni big data technológiák fejlesztésével, ezek nagy része pedig Open Source keretek közé került vagy már eleve ott is jött létre. Ezek egy része magasabb szintű, programozás nélkül is alkalmazható adatkezelési, elemzési lehetőségeket ad Hadoop felhasználóknak (pl. Hive, Impala, egyéb SQL on Hadoop megoldások), más része pedig olyan technológiák csoportja, amelyek nem HDFS alapúak, de szintén big data technológiák, amelyek a Hadoop korlátait kerülik ki. Adatbázis alapon történő tárolás (pl. Greenplum Database), memóriában történő feldolgozás (Apache Spark), logok és egyéb információk valós időben történő továbbítása, replikációja (Flume, Kafka). Egy részük pedig a Hadoop környezet menedzsment-jét vagy az adattöltési és feldolgozási folyamatok támogatását szolgálja (pl. Apache Yarn, Zookeeper).

Ugyanakkor a különböző helyeken keletkező, egyre nagyobb mennyiségű adat kezelése és elemzése kapcsán előtérbe kerültek olyan korábban már létező technológiák is, amelyek kitörtek az addigi szűkebb ismertségből és egyre elterjedtebbé váltak (Key-Value stores, NoSQL adatbázisok, Event Processing, Machine Learning). Számos új cég jelent meg új termékekkel, illetve számos Open Source projekt indult el vagy vált jelentősebbé (pl. Open Source R).

Az open source gyökerek és a nagyon sok kisebb-nagyobb cég vagy fejlesztő csoport által fejlesztett különböző szintű, terjedelmű és funkcionalitású szoftverek mennyisége miatt a big data-hoz társult (és bizonyos fokig még ma is társul) egyfajta „csináld magad” szemlélet, ahol is a felhasználók saját maguk válogatják össze a különböző hardver és szoftver komponenseket, és integrálják őket kész megoldássá. Ezért már korán megjelentek azok a cégek, amelyek integrált big data megoldásokat kínálnak a felhasználóknak szoftver szinten (pl. Pivotal, Cloudera, Hortonworks) vagy hardver szinten (NetApp, EMC).

Az egyre fejlettebb megoldások megjelenésével már nem csak az internetes cégek és egzotikus startup vál-



1. ábra Google kereső-kifejezések statisztikája

lalatok érdeklődtek a big data iránt, hanem megjelentek a hagyományosabb piaci szegmensekben lévő nagy kereskedelmi, pénzügyi és termelő vagy kutató vállalatok is, akiknél a fő hajtóerő az eddig kihasználatlanul tárolt (vagy még csak nem is tárolt) adatvagyon elemzésével elérhető versenyelőny volt. Ezen cégek jellemzően a legnagyobb IT vállalatok megoldásait használják, így hamarosan a legnagyobb informatikai cégek is elkezdték kifejleszteni a saját megoldásaikat (Oracle, IBM, SAP, Amazon Web Services), amelyek kifejlesztésekor a hardver és szoftver szintű integráció mellett fontos szempont a más szoftvereknél már megszokott, és nagyvállalati környezetben fokozottan igényelt egyéb funkcionalitások biztosítása volt: biztonság, menedzselhetőség, integrálhatóság.

2. A big data technológia alkalmazása

A big data technológiák tehát egyre szélesebb körben kezdtek elterjedni, alternatívát kínálva a hagyományos relációs adatbázisokon alapuló adattároláshoz képest. A főbb motivációs tényezők, amelyek ezt a folyamatot gerjesztik az alábbiak:

- *Költséghatékonyság:*
  A relációs adatbázis és SAN alapú tárolás infrastrukturális költségénél alacsonyabb költséggel lehet tárolni az adatokat.
- *Adatbetöltési teljesítmény:*
  Nagy sebességgel, nagy mennyiségben keletkező adatok tárolásánál a big data technológiák alkalmazása előnyösebb lehet a hagyományos adatbázis alapú tároláshoz képest.
- *Lekérdezési/elemzési teljesítmény:*
  Olyan elemzési célú lekérdezéseknél, amelyek egyszerre nagy adatmennyiséget mozgatnak meg, akár többször is, a big data technológiák gyorsabbak lehetnek egy relációs adatbázishoz képest.
- *Valós idejű adatfeldolgozás:*
  Nagy sebességgel, nagy mennnyiségben keletkező adatokon történő azonnali transzformációk elvégzése hatékonyabb lehet.
- *Struktúrálatlan vagy lazán struktúrált adatok tárolása, elemzése:*
  Ilyen típusú adatok tárolása és lekérdezése relációs adatbázis környezetben nehézkesebb vagy nem feltétlenül a legjobb választás.

A költséghatékonysággal és a teljesítménnyel kapcsolatban fontos megjegyezni, hogy nincsenek kvantitatív módon megfogalmazható szabályok arra, mikor érdemes big data technológiát használni. A relációs adatbázisokhoz képest elérhető költség- és teljesítményelőny csak az alternatíva költségével és teljesítményével együtt értelmezhető. A mai modern adatbáziskezelők számos olyan funkcionalitást tudnak nyújtani, amelyek nagy tömegű adat gyorsabb, hatékonyabb kezelését biztosítják relációs környezeten belül. Ehhez felhasználnak szoftver megoldásokat (melyek használata addicionális költséget jelenthet), igényelhetnek fokozott mértékű redundanciát és sebességet biztosító (drágább) tároló technológiákat, illetve speciális céleszközöket, ahol hardver és szoftver technológiák együttes működése biztosítja a kívánt teljesítményt. Számos cégnél – különösen a nagyvállalatoknál – olyan relációs adatbázis infrastruktúra van, amelyben rutinszerűen kezelnek több száz terabyte-nyi adatot. Amennyiben az infrastruktúra mellett a kapcsolódó adatbetöltéshez és utána az elemzéshez szükséges kapacitás is rendelkezésre áll, akkor nincs ok a meglévő infrastruktúra mellé egy új, big data infrastruktúrát is kialakítani. Általában az infrastruktúra rendelkezése állása (szoftver technológia, tárkapacitás) a kevésbé problémás és a betöltéshez, feldolgozáshoz szükséges elemzési kapacitás („gépidő”) az, amit vagy alábecsülnek, vagy nem kalkulálnak vele kellő mértékben, ami egyrészt a betöltési és feldolgozási időablakok elhúzódásával, másrészt az elemzők felé elérhető teljesítmény korlátozásával jár. Ez akár a tervezett elemzési felhasználást is ellehetetlenítheti.

A költség és technológiai korlátok mellett, az elemzési lehetőségekben rejlő előnyök azok, amelyek okán a big data technológia alkalmazása elkezdődhet. Erre egy példa lehet a telekommunikációs vállalatoknál történő lemorzsolódás elemzése. Egy ilyen elemzésben azt próbáljuk megjósolni minden ügyfelünkről különkülön, hogy ügyfelünk marad-e vagy pedig a közeljövőben várható, hogy felmondja a szerződését (és valószínűleg egy versenytársunkkal köt új szerződést.) Az elemzés kimenetele ügyfelenként egy boolean típusú változó, amely jelzi, hogy várhatóan lemorzsolódik-e az ügyfél vagy nem. Egy ilyen elemzésben kétféleképpen is tévedhetünk: lemorzsolódónak jósolunk valakit, aki nem fog lemorzsolódni (hamis pozitív) vagy nem jósolunk lemorzsolódónak valakit, aki pedig el fog hagyni minket (hamis negatív).

Maga az elemzés úgy történik, hogy veszünk egy kellően nagy mintát a korábban már lemorzsolódott ügyfeleinkből, egy kontrollcsoportot a nem lemorzsolódott ügyfelek közül, és ezekhez az ügyfelekhez összegyűjtjük azok rendelkezésre álló adataikat: demográfiai adatokat, feléjük kiállított számlák adatait, fizetési tranzakcióik adatait, hívás statisztikai adataikat, illetve egyéb olyan adataikat, amelyek a vállalat birtokában vannak az ügyfelekről. Az összegyűjtött adatokon aztán olyan adatbányászati algoritmusokat futtatunk, amelyekkel összefüggést keresünk az ügyféladatok és a lemorzsolódás ténye között. Ez egy többlépcsős elemzési folyamat, amelynek során egyre több adatot vonunk be az elemzésbe, a forrásadatainkon különböző statisztikai vagy más transzformációkat végzünk és elemezzük az eredményeket.

Egy sikeres elemzés végén a rendelkezésünkre fog állni egy olyan algoritmus, ami az elemzésbe bevont jelenlegi ügyfél adatok alapján viszonylag jó hatásfokkal megjósolja azt, hogy egy ügyfél várhatóan elhagye minket a közeljövőben vagy veszélyeztetett-e ilyen szempontból. Az algoritmus hatékonysága aztán a későbbi időszak tényadatainak függvényében értékelhető, elemezhető. Az ilyen elemzések az mutatják, hogy

minél többféle adatot vonunk be az elemzésbe, annál pontosabban tudjuk megjósolni a végeredményt.

A *2. ábrán* látható grafikon X tengelyén a hamis negatív találatok, Y tengelyén az igaz pozitív találatok szerepelnek. Attól függően, hogyan kalibráljuk az algoritmusunkat, egyre több igazi pozitív találatunk van, de ugyanakkor egyre több lesz a hamis negatív találatunk is. Egy görbe azt szemlélteti, hogy a paraméterezés függvényében hogyan alakul az algoritmusunk jóslási pontossága. Az ideáli pont a bal felső sarokban lenne, ahol is csak igaz pozitív találataink vannak, és nincsenek hamis negatív találatok. A való életben persze ezt nem lehet elérni, de az ábra jól mutatja azt, hogy ha egyre többféle ügyféladatot vonunk be az elemzésbe, akkor egyre pontosabb elemzést tudunk készíteni.

A hívásadatok elemzésben történő felhasználásához már szükség lehet big data megoldásra is. A legtöbb telekommunikációs cégnél a hívásadatok rendelkezésre állnak valamilyen szűkített formában relációs adatbáziskezelőben. A szűkítésnek számos oka lehet, ugyanakkor elemzési korlátokat okozhat. Például ha csak azok a hívásrekordok állnak rendelkezésre, amelyeknek pénzügyi vonzata van a vállalat szempontjából, akkor hiányzik az az információ, hogy az ügyfél csak megcsörgetett egy telefonszámot (amiről aztán visszahívták azonnal). Ha a pénzügyi vonzattal rendelkező hívásrekordról hiányzik az az információ, hogy miért fejeződött be a hívás (pl. hálózati hiba miatt megszakadt), akkor ismét egy elemzés szempontjából fontos információ veszett el. Az ilyen – első látásra üzleti szempontból nem kulcsfontosságú – információk miatt lehet hasznos a hívásadatok tárolása relációs adatbázis mellett big data technológiával is.

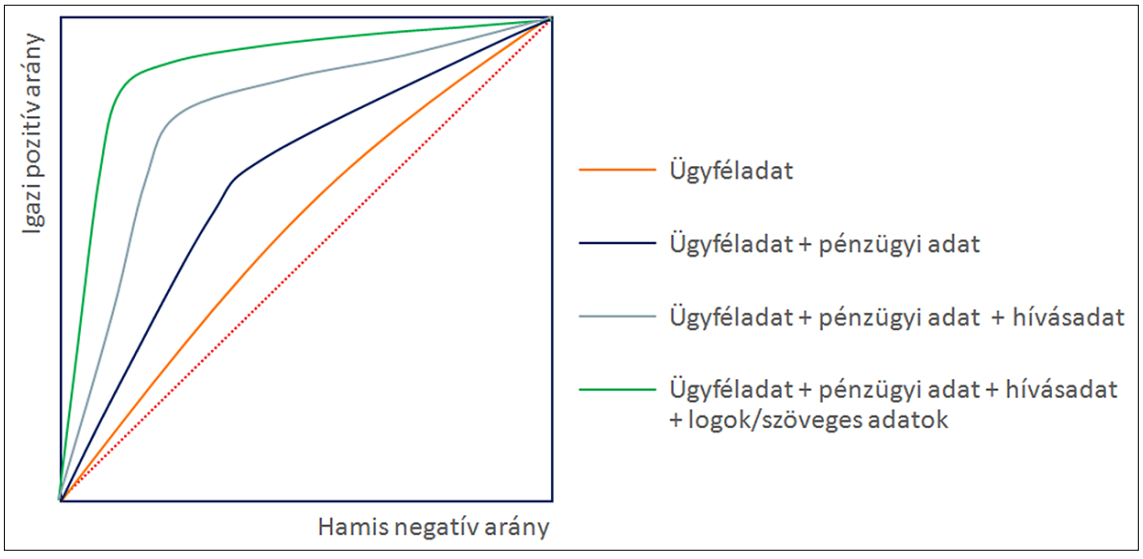
A big data segítségével pedig még tovább terjeszthetjük ki az elemzésünk pontosságát olyan adatokkal amelyek legtöbbször semmilyen formában nem érhetők el az üzleti elemzők számára: ügyfélpanaszok szöveges formában, hálózati adatforgalmi logok, amelyekből következtethetünk az ügyfél felhasználási élményére: sebesség, megszakadt adatkapcsolatok száma, gyakorisága stb. De ha közösségi médián keresztül is kapcsolatban vagyunk az ügyféllel, akkor esetleg láthatjuk a termékeinkről szolgáltatásokról alkotott véleményét is: ajánlotta-e másoknak a szolgáltatásainkat, panaszkodott-e nyilvános fórumon stb.

3. Az Oracle big data koncepciója

Az Oracle szerint a big data megoldások nem helyettesíthetik a hagyományos relációs adattárolási stratégiákat, viszont a vállalati adatvagyon egészét tekintve nagyon sok esetben lehet létjogosultsága a big data technológiáknak. Alkalmazásuk egyik akadálya az, hogy a megfelelő szakértelem nem áll rendelkezésre vállalaton belül: a big data infrastruktúra felépítéséhez speciális szakértelem szükséges. Amennyiben saját magunk akarjuk felépíteni a teljes infrastruktúrát számos hardver és szoftver gyártó termékei közül kell mérlegelnünk, kiválasztanunk a számunkra leginkább alkalmas termékeket, amelyek együtt tudnak működni egymással.

A hardver cluster felépítése és konfigurálása (beleértve a cluster hálózati kapcsolatait is), valamint a rajta futó szoftverek konfigurálása, végül az egész rendszer teljesítmény hangolása akár hónapokat is igénybe vehet, mielőtt egyáltalán elkezdődhetne az éles adatbetöltés vagy adatelemzés. Sok teljesítmény hangolási vagy konfigurálási probléma csak a tényleges használatba vétel után derül ki.

Ezért fejlesztette ki az Oracle speciálisan a big data megoldások számára az Oracle Big Data Appliance nevű termékét, amely egy ügyfél igényeinek megfelelően méretezett, installált és konfigurált saját tárolóval rendelkező szerver-cluster, amely néhány napos üzembehelyezés után egy produktív, üzemszerűen működtethető komplett big data környezet. A hagyományos big data technológiákon kívül tartalmaz olyan, a nagyvállalatok számára különösen fontos biztonsági és management megoldásokat, amelyeket más környezetekben már standard-nek számítanak, de az Open Source big data megoldások világában még nem: Kerberos alapú autentikáció, LDAP alapú autorizáció, Oracle En-



2. ábra Lemorzsolódás elemzés hatékonysága

terprise Manager Cloud Control, Oracle Audit Vault és Database Firewall.

Fontos, hogy az ilyen módon tárolt adatok ne szigetszerűen létezzenek, hanem összekapcsolhatóak és integrálhatóak legyenek a hagyományos formában tárolt adatokkal. Az integrációnak és az adatok összekapcsolhatóságának számos vetülete létezik:

- Eseményfeldolgozás: a nagymennyiségű, nagy sebességgel érkező adat on-the-fly feldolgozása, majd perzisztens tárolása relációs vagy big data környezetben (Oracle Complex Event Processing)
- On-line szinkornizáció big data és relációs rendszerek között (Oracle GoldenGate)

* Batch alapú adatmozgatás és/vagy komplex transzformációkat végző adattáttöltések big data és relációs rendszerek között (Oracle Data Integrator)
* Komplex ad-hoc lekérdezések futtatása rendszereken keresztül (Oracle Big Data SQL). Segítségével az elemzők Oracle SQL lekérdezéseket futtathatnak big data adattárolóban tárolt adatokon, akár közvetlenül összekapcsolva a relációs adatbázisokban tárolt adatokkal.

Ezek a szoftver megoldások mind arra szolgálnak, hogy átjárhatóságot biztosítsanak a big data, a hagyományos relációs világ és egyéb más adatbázisok vagy rendszerek között. Ehhez kapcsolódik az Oracle Konzultáció által biztosított szakértelem, amely a mély termékismereten túl sok komoly bevezetési projekt tapasztalatán alapuló informatikai tanácsadással tud hozzájárulni egy sikeres big data bevezetési projekthez.

4. Összefoglalás

A big data megoldások elterjedése egyre inkább jellemző lesz minden iparágban és az állami szektorban egyaránt. A kezdeti „úttörő” felhasználók után egyre többen fognak törekedni arra, hogy az eddiginél magasabb szinten aknázzák ki meglévő adatvagyonukat vagy versenyelőnyre tegyenek szert. A számos különböző termék és megvalósítási alternatíva előtt álló felhasználóknak a konkrét elérendő célok mellett arra is érdemes figyelni, hogy a bevezetendő új alkalmazások minél szorosabban tudjanak illeszkedni a meglévő rendszerekhez és az alkalmazott informatikai szabványokhoz. Előny lehet egy bevezetési projektben, ha olyan szállítóhoz tudnak fordulni, amely átfogó felelősséget tud vállalni az összes hardver és szoftver komponens működéséért valamint a teljes rendszer bevezetéséért is.

A szerzőről

STADLER GELLÉRT 1996-ban szerzett diplomát az Egri Eszterházy Károly Főiskolán. 2007-től az IBM Magyarország Kft. rendszerintegrációs részlegén kezdett dolgozni. Elsősorban adattárház és üzleti intelligencia rendszerek tervezésével és fejlesztésével foglalkozott. 2005-től az Oracle Hungary Kft. tanácsadójaként dolgozik az adattárház és BI csoportban. E szerepeiben több magyarországi nagyvállalatnál is végzett tanácsadói munkát: Bricostore Hungária Kft., AUDI Hungária Zrt., ING Biztosító ZRt., Vodafone Magyarország, Budapest Airport Zrt., Generali Zrt., FHB Bank Zrt., Budapest Bank Zrt.

Irodalom

[1] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung, „The Google File System”, October 2003. *http:///labs.google.com/papers/gfs.html*

[2] White, Hadoop: The Definitive Guide, 3rd Edition, 2012. május 19., O’Reilly Media, Inc.

[3] Jeffrey Dean, Sanjay Ghemawat, „MapReduce: Simplified Data Processing on Large Clusters”, December 2004. *http:///labs.google.com/papers/mapreduce.html*