A FEJEZET TARTALMÁBÓL:

- Az adatok nyomába szegődünk
- Fontos adatforrások áttekintése
- Mi a különbség az adatelemzés és az adattechnika között?
- Adatok tárolása saját szerveren vagy a felhőben
- További adatmérnöki megoldások

2. fejezet
Az adattechnika

Bár az adatok és a mesterséges intelligencia (MI) rendkívül érdekes témák a nagyközönség szemében, a legtöbb laikus nem tudja, hogy valójában mik is azok az adatok, és hogyan használják őket ahhoz, hogy jobbá tegyék az emberek életét.

Ebben a fejezetben a modern adat-ökoszisztémák teljes történetét olvashatod – annak a magyarázatát, hogy honnan származnak az adatok és hogyan használják őket, valamint, hogy mi a gépi tanulási mérnökök, az adatmérnökök és az adatelemzők szerepe ebben a folyamatban. Ebben a fejezetben bemutatjuk az adatok tárolásának és feldolgozásának adatelemzéssel kapcsolatos alapvető fogalmait, így ezek a tudnivalók alapul szolgálhatnak ahhoz, hogy kidolgozd a terveidet az üzleti teljesítmény adatelemzéssel való javításához.

Mi az a három V?

Ha a vállalatok versenyképesek akarnak maradni, akkor ügyesen és hozzáértő módon kell tudniuk beépíteni az adatokból nyert mögöttes információkat a folyamataikba és termékeik-

be, valamint növekedési és menedzsment stratégiáikba. Ez fokozottan igaz a digitális átállásra, amely a COVID–19 világjárvány közvetlen következményeként rendkívüli ütemben felgyorsult. Akár terabájtos, akár petabájtos nagyságrendű az adatmennyiséged, az adattechnikai alapú megoldásokat úgy kell megtervezni, hogy azok megfeleljenek az adatok rendeltetési helyére és felhasználására vonatkozó követelményeknek.

Három jellemző a leginkább meghatározó, melyek alapján értelmezheted az adataidat: a mennyiség, a sebesség és a sokféleség – ezeket nevezik „a három V-nek" az angol elnevezésük (volume, velocity, variety) alapján. Mivel az adatok ezen három jellemzője folyamatosan nő, folyamatosan új, egyre innovatívabb technológiákat kell kifejleszteni ezeknek a problémáknak a kezelésére.

Megküzdés az adatmennyiséggel

Nyers formában a legtöbb adat *alacsony értékű* – más szóval, a nyers adatokban az érték/adatmennyiség arány alacsony. Sok adat hatalmas számú, de apró, különböző formátumban érkező tranzakcióból áll. Ezek az adatelemek csak akkor jelentenek valódi értéket, ha összesítik és elemzik őket. Leegyszerűsítve azt mondhatjuk, hogy az adatmérnökök feladata az adatok összesítése, az adatelemzők feladata pedig azok elemzése.

Az adatsebesség kezelése

A legtöbb adatot automatizált folyamatok és különböző eszközök hozzák létre, és mivel az adattárolás költségei viszonylag alacsonyak, sokszor a rendszer sebessége a korlátozó tényező. Ne feledd, hogy a nyers adatok értéke alacsony. Ebből következik, hogy olyan rendszerekre van szükség, amelyek rövid idő alatt nagy mennyiségű adatot képesek feldolgozni, hogy aktuális és értékes mögöttes információkat nyerjenek ki.

Definició szerint az *adatsebesség* az adatok mennyisége egységnyi idő alatt. Minden adatrendszer jellemzője a késlelte-

tés, amely azt a késést számszerűsíti, amellyel a rendszer az adatokat mozgatja, miután erre utasítást kapott. Sok adattechnikai rendszerrel szemben elvárás, hogy 100 ezredmásodpercnél is kisebb legyen a késleltetése, az adatok létrehozásának időpontjától a rendszer válaszáig mérve.

Az *átviteli sebesség* egy olyan jellemző, amely azt írja le, hogy a rendszer egységnyi idő alatt mennyi munkát képes elvégezni. Az adatrendszerekben az elvárt átviteli sebesség könnyen elérheti az 1000 üzenet/másodpercet is! A nagy sebességű, valós időben mozgó adatok akadályozzák az időben történő döntéshozatalt. Az adatkezelési és adatfeldolgozási technológiák képességei gyakran korlátozzák az adatátviteli sebességet.

Az eszközöknek, amelyek adatokat juttatnak a rendszerbe – más néven az adatbeviteli eszközöknek – számos különböző fajtája van.

Az adatok sokféleségének kezelése

Még bonyolultabb lesz értelmezni az adatokat, ha strukturálatlan és félig strukturált adatokat is hozzáadunk a strukturált adatforrásokhoz. Ez a *rendkívül* sokféle adat számos forrásból származik. A legszembetűnőbb jellemzőjük, hogy különböző alapszerkezetű (strukturált, strukturálatlan vagy félig strukturált) adathalmazok kombinációjából állnak. A heterogén, nagy sokféleségű adatok gyakran gráfadatok, JSON-fájlok, XML-fájlok, közösségimédia-adatok, strukturált táblázatos adatok, webes naplóadatok és weboldalakon végzett felhasználói kattintásokból generált adatok – más néven *kattintássorozatok* – tetszőleges kombinációiból állnak.

A *strukturált adatok* relációs adatbázis-kezelő rendszerben (Relational database management system, RDBMS) tárolhatók, feldolgozhatók és kezelhetők – ilyen típusú rendszer például egy PostgreSQL-adatbázis, amely sorokból és oszlopokból álló táblázatos sémát használ, megkönnyítve ezzel a konkrét értékek azonosítását az adatbázisban tárolt adatok között. Ezek az adatok, amelyeket emberek vagy gépek is generálhatnak, mindenféle forrásból származhatnak – a kattintássoroza-