A Big Data elemzési folyamat kritikus fázisai

Élő Gábor[1] – Szármes Péter[1,2]

[1] Széchenyi István Egyetem, Információs Társadalom Oktató- és Kutatócsoport
[2] Széchenyi István Egyetem, Multidiszciplináris Műszaki Tudományi Doktori Iskola

A Big Data elemzések analitikai és gépi tanuló technikák olyan kombinációját jelentik, amelyekkel rejtett mintázatokat, összefüggéseket lehet felderíteni nagy adathalmazokban. A módszerek közé tartozik a regresszióanalízis, az asszociációs szabályok, az optimalizáció, a Monte Carlo szimuláció, stb. Ezekkel a módszerekkel összetettebb kérdéseket lehet megválaszolni, és jelentős értéket lehet teremteni az adott szervezet számára.

Sok elemzési probléma nagynak és ijesztőnek tűnik elsőre, de egy jól definiált analitikai folyamat segít a komplex problémák kezelhető feladatokra történő lebontásában. A jól felépített analitikai folyamat áttekinthető, megismételhető módszert ad az elemzés elvégzésére. Segít az idő megfelelő beosztásában, például a folyamat elején kellő figyelmet fordít az üzleti probléma világos megfogalmazására. A folyamat menetének rögzítése és dokumentálása kifejezi az eredmények megbízhatóságát is, és nagyobb hitelt ad az elemzési projektnek. Emellett elősegíti a módszerek átadását is, hogy azok megfelelően megismételhetők legyenek a jövőben. Egy nagyobb projektben számos különböző szereplő vesz részt, ezért rendkívül fontos a munka összehangolása, amiben szintén nagy segítséget ad egy gondosan kidolgozott folyamat.

Kulcsszavak: big data, üzleti analitika, elemzési folyamat

Critical phases of the Big Data analytical process

Big Data analytics is a combination of analytical and machine learning techniques helping to discover hidden patterns and interactions in very large and diverse datasets. The methods include regression analysis, association rules, optimization, Monte Carlo simulation, etc. With these methods you can find the answer to complex questions and create significant value for your organization.

Many analytical problems seem big and intimidating at first, but a well-defined analytical process can help to break a complex problem into manageable tasks. A structured analytical process gives a transparent and repeatable method for conducting complex analyses. Describing and documenting all the steps in the process increases the reliability of results and generates a bigger confidence for the analytical project. It also facilitates the transmission of methods in order to repeat the process when solving similar problems in the future. In a large project with several different stakeholders the coordination of their work is of crucial importance, and a well-defined process can help here as well.

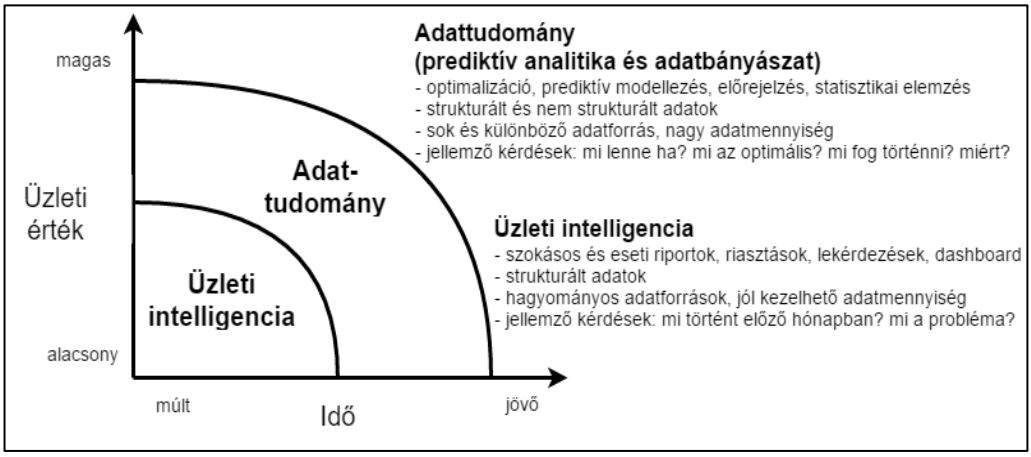
Keywords: big data, business analytics, analytical process

1. BEVEZETÉS

Az üzleti intelligencia (BI) egy konzisztens metrikára fókuszál, amellyel mérhető a vállalat múltbeli teljesítménye több dimenzióban (például Balanced Scorecard rendszer), és amely alapjául szolgál az üzleti tervezésnek. Ide tartozik a teljesítménymutatók kialakítása, amelyeken keresztül követhető aztán a vállalat teljesítménye. A mérőszámokat és indikátorokat általában egy OLAP[14] rendszerbe töltik be, és ez szolgál a különféle riportok alapjául. A prediktív analitika és az adatbányászat (data science) ezzel szemben analitikai és gépi tanuló technikák olyan kombinációjára utal, amelyekkel rejtett mintázatokat, összefüggéseket lehet felderíteni az adatokban. A módszerek közé tartozik a regresszióanalízis, az asszociációs szabályok, az optimalizáció, a Monte Carlo szimuláció, stb. Ezekkel a módszerekkel összetettebb kérdéseket lehet megválaszolni, és komoly értéket lehet teremteni a szervezet számára.

Mind a Business Intelligence mind a prediktív analitika, adatbányászat szükséges a vállalat működéséhez, az üzleti kihívások sikeres megválaszolásához. Az adatelemzési tudomány azonban gyakran Big Data feladatokkal, nagy mennyiségű, hiányos vagy nem strukturált adatokkal foglalkozik, ami nagyobb szorgalmat, adatelőkészítést és adattisztítást követel meg, mint egy hagyományos Business Intelligence projekt, amely jellemzően jól strukturált adatokkal dolgozik egy adattárházban vagy OLAP kockában (1. ábra). A Big Data analitikai feladatoknál ezért még fontosabb a strukturált munkavégzés, hogy a projekt valóban elérje a kitűzött célokat.

1. ábra Az üzleti intelligencia és az adattudomány viszonya



Forrás: EMC 2013.

[14] OLAP (Online Analytical Processing): adatok interaktív elemzését szolgáló rendszer, amelynek lényege egy többdimenziós adatmodell, amely lehetővé teszi komplex lekérdezések gyors végrehajtását (Codd és munkatársai 1993).

2. AJÁNLOTT ELEMZÉSI FOLYAMATOK

Sok elemzési probléma nagynak és ijesztőnek tűnik elsőre, de egy jól definiált analitikai folyamat segít a komplex problémák kezelhető feladatokra történő lebontásában. A jól felépített analitikai folyamat áttekinthető és megismételhető módszert ad az elemzés elvégzésére. Segít az idő megfelelő beosztásában, például a folyamat elején kellő figyelmet fordít az üzleti probléma világos megfogalmazására. Gyakori hiba, hogy az adatgyűjtés és –elemzés elkapkodott elkezdése miatt, nem fordítanak elegendő időt az üzleti probléma körüljárására. Ennek a következménye az lehet, hogy a projekt közepén a résztvevők azt veszik észre, hogy az üzleti szponzorok más probléma megoldását keresik, mint az elemzők. A folyamat menetének rögzítése és dokumentálása kifejezi az eredmények megbízhatóságát is, és nagyobb hitelt ad a projektnek. Emellett elősegíti a módszerek átadását is, hogy azok megfelelően megismételhetők legyenek a következő negyedévben, a következő évben, vagy az új munkatársak könnyebben betanuljanak a munkába.

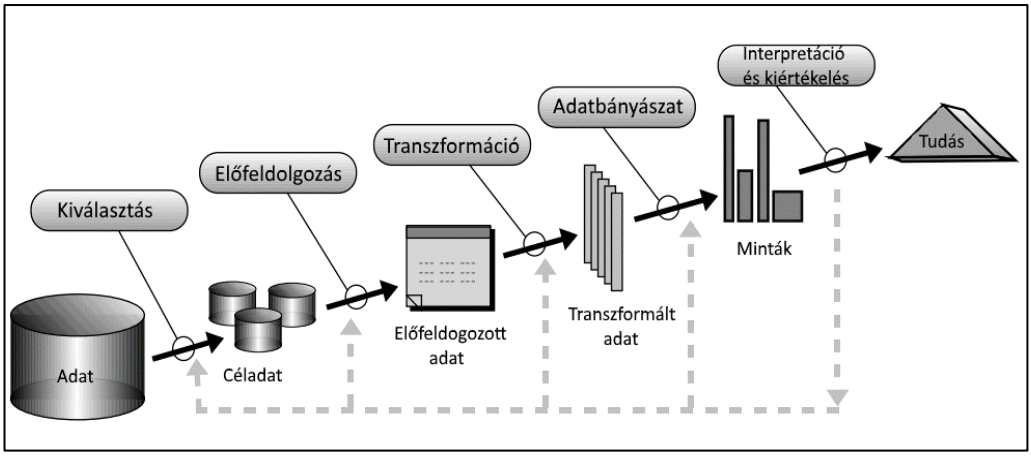
Az ismertetett elemzési folyamatok elsősorban komplex elemzési feladatokra vonatkoznak nagyobb, üzleti tevékenységet végző szervezeteknél. A megállapítások azonban részben igazak kisebb cégekre és tudományos vagy egyéb tevékenységet végző szervezetekre is, ha összetett elemzési tevékenységet végeznek nagy mennyiségű és különböző fajta adatokkal. A hasonló feladatok sikeres menedzseléséhez szinte elengedhetetlen egy jól átgondolt és dokumentált folyamat, amelyet természetesen az adott tevékenység és szervezet igényeihez, sajátosságaihoz kell igazítani.

Fayyad és szerzőtársai elméletibb, általános jellegű ajánlásokat adnak, amelyek lényegében ma is érvényesek, bár az 1990-es évek közepén fogalmazták meg őket. A Forrester és az EMC javaslatai a közelmúltból származnak, és jóval gyakorlatiasabb szemléletűek, az elemzési folyamat felépítésénél figyelembe veszik a manapság alkalmazott informatikai rendszerek jellemzőit is. Az EMC tanulmánya pedig ezen felül figyelembe veszi a (nagy)vállalati környezet sajátosságait is: kitér a gazdaságosság és a megtérülés kérdésére, a vállalati kommunikációs fontosságára, stb.

2.1. TUDÁSFELTÁRÁS ADATBÁZISOKBAN

Fayyad és szerzőtársai klasszikus cikkükben (Fayyad et al. 1996) leírják az adatbázisokban rejlő tudás felfedezésének (Knowledge Discovery in Databases, KDD) folyamatát. A tudás felfedezésének folyamata erősen iteratív, és bármelyik két lépés között lehetnek hurkok. A lépések alapvető lefutását (az iterációk figyelembe vétele nélkül) a 2. ábra illusztrálja.

2. ábra Az adatbázisokban rejlő tudás felfedezési folyamata



Forrás: Fayyad et al. 1996.

A KDD folyamat interaktív és iteratív, több lépésből épül fel:

Az első lépés az adott alkalmazási terület megértése és a felfedezési folyamat céljához szükséges előzetes tudás megszerzése. Ehhez meg kell ismerni az ügyfél szempontrendszerét is.

A második lépés a cél adathalmaz létrehozása: a rendelkezésre álló adatokból ki kell választani azokat az adathalmazokat, illetve azokat a változókat vagy mintákat, amin aztán elvégzik a tényleges elemzéseket.

A harmadik lépés az adatok tisztítása és előzetes feldolgozása. Az alapvető műveletek közé tartozik a zaj eltávolítása, döntés a hiányzó adatok kezeléséről, az időbeli változások felmérése.

A negyedik lépés az adatok redukciója és projekciója: ki kell választani az adathalmazt jól reprezentáló jellemzőket a feladat céljától függően. A dimenzionalitás csökkentésével vagy különböző transzformációs módszerekkel, a vizsgált változók száma csökkenthető, ami elősegíti az elemzést.

Az ötödik lépés a tudásfeltárási folyamat első pontban meghatározott célját segítő adatbányászati módszerek kiválasztása. Például az adathalmaz vizsgálata statisztikai módszerekkel, illetve klasszifikáció, regresszió, klaszteranalízis, stb.

A hatodik lépés az adatok feltáró elemzése, a modellek és hipotézisek kiválasztása. Ide tartozik a megfelelő adatbányász módszerek és algoritmusok kiválasztása, amelyek segítségével feltárhatók az adatokban rejtőző minták. Ide tartozik a modellekről és paramétereikről történő döntés: másképp kell kezelni egy kategorikus adatokra vonatkozó modellt, mint egy valós változók vektoraira vonatkozó modellt. Figyelembe kell venni az ügyfél szempontjait, aki számára például fontosabb lehet, hogy értse a modell működését, mint a modell előrejelző képessége.

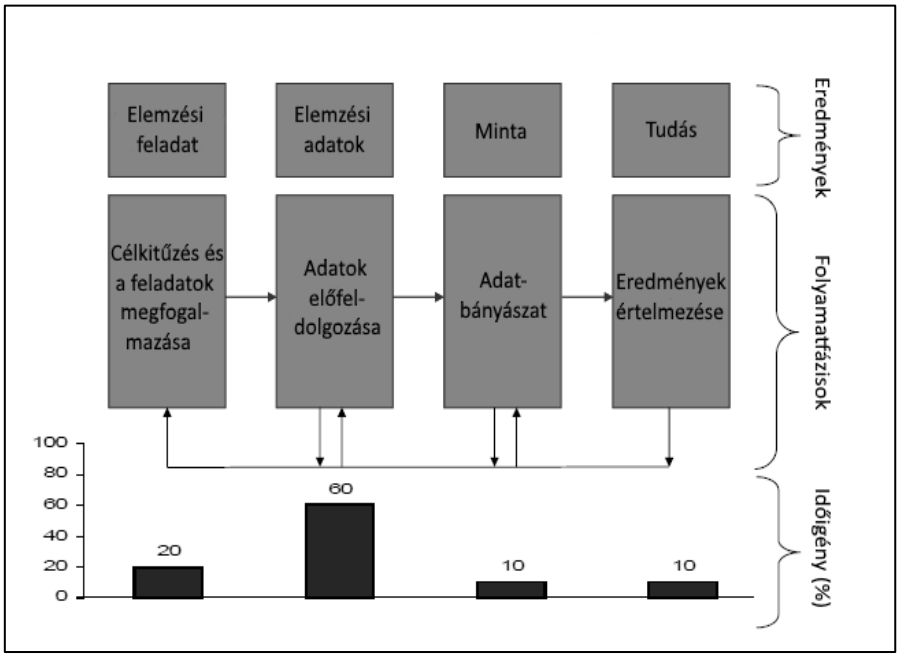
A hetedik lépés az adatbányászat: érdekes minták keresése az adatokban klasszifikációs szabályokkal, döntési fákkal, regresszióval, klaszteranalízissel. Az előző lépések gondos elvégzése nagyban elősegíti ennek a lépésnek az eredményességét.

A nyolcadik lépés a feltárt minták értelmezése, vagy az előző lépések iteratív megismétlése. Ide tartozhat a feltárt minták, modellek, illetve az adatok megfelelő vizualizációja is.

A kilencedik lépés a feltárt tudásnak megfelelő cselekvés a tudás direkt felhasználása révén vagy a kinyert információ bevitele egy másik rendszerbe, vagy akár egyszerű dokumentációja és kommunikációja a megfelelő felek irányába. Ezen a ponton kell megvizsgálni a feltárt tudáselemek és a már meglévő (vélt vagy feltárt) tudáselemek közötti konfliktusokat.

A tudásfeltárással foglalkozó munkák többsége a hetedik lépésre az adatbányászatra fókuszált. A folyamat gyakorlati sikeréhez azonban a többi lépés is annyira vagy még inkább fontos: például csak az adatok előkészítése a teljes folyamat időigényének kb. 60%-t teszi ki (3. ábra).

3. *ábra* A folyamat szakaszainak időigénye



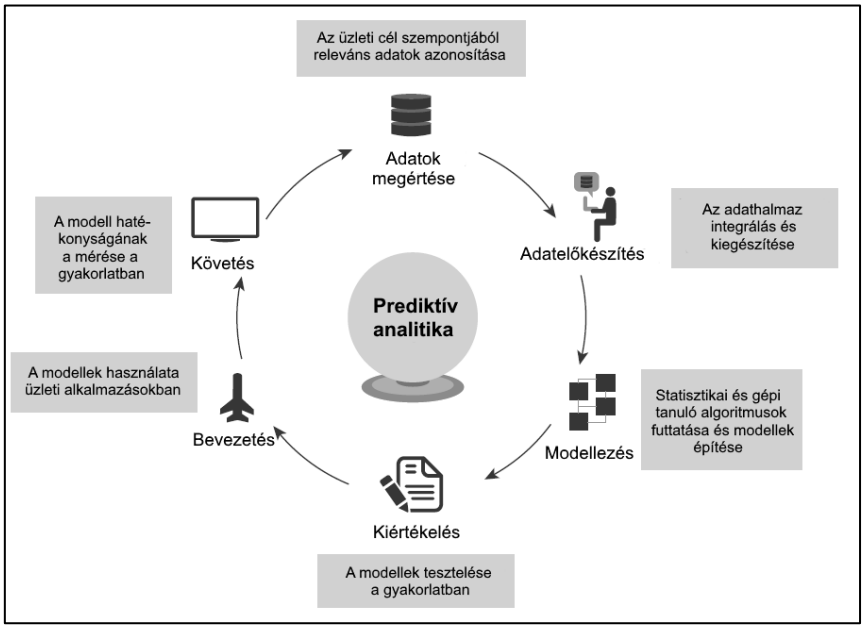
Forrás: Felden 2015.

2.2. A FORRESTER ÁLTAL AJÁNLOTT ELEMZÉSI FOLYAMAT

A Forrester Research technológiai piackutató vállalat elemzése a prediktív analitikai megoldásokról hasonló gondolatokat fogalmaz meg, mint Fayyad és szerzőtársai klasszikus cikke (Gualtieri és Curran 2015). A jó elemzési folyamat mindig a megfelelő kérdésekkel kezdődik. Üzleti környezetben például valamilyen folyamat optimalizálása lehet a cél, és a kérdés pedig az lehet, hogy hogyan lehet növelni az értékesítést, az árakat, a profitabilitást vagy a hatékonyságot. Gyakori cél valamilyen kockázat csökkentése, itt olyan kérdések merülhetnek fel, hogy mely ügyfelek készülnek elpártolni a cégtől, melyik ügyfél nem tudja majd várhatóan fizetni a törleszést (például egy banknál), vagy mely ügyfeleknél valószínű visszaélés (például egy biztosítónál).

A prediktív analitika különböző algoritmusok segítségével mintákat keres az adatokban, amelyek alapján előre jelezhetők jövőbeli események: például megalkotható egy olyan modell, amely megjósolja, hogy várhatóan mely ügyfelek fordulnak el a cégtől. Egy telekommunikációs cég bizonyos vevőadatok (például a hívások száma, hossza, SMS üzenetek száma, havi átlagos számlaérték és sok más változó) alapján alkothat egy modellt, ami meghatározza, hogy várhatóan mely ügyfelek váltanak át más mobilszolgáltatóhoz. Ha cég az elemzés segítségével meg tudja határozni, hogy mi az oka az ügyfelek átvándorlásának, akkor még időben lépéseket tehet, hogy ezt megelőzze. Ez az elemzési folyamat nem egyszeri eset, a cégnek az új, aktuális adatokon újra és újra le kell futtatni az elemzést, hogy a modellek érvényesek legyenek, tükrözzék az aktuális piaci helyzetet. Egyes cégek hetente végzik az elemzést, mások lényegében folyamatosan.

4. ábra Az adatokban rejlő tudás kinyerésének lépései



Forrás: Gualtieri és Curran 2015.

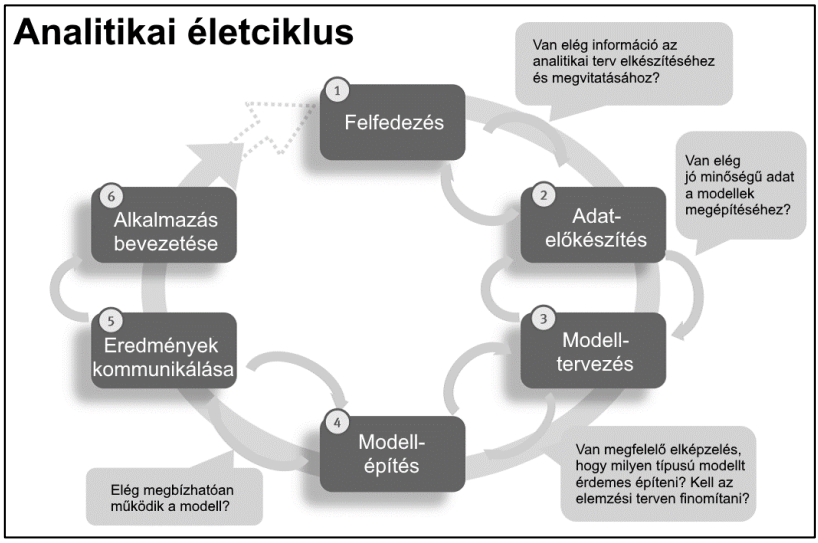
Az igazi felismerések mindig mély, kreatív kérdésekkel kezdődnek. Ha megvannak a kérdések, akkor a Forrester a következő hat lépést javasolja a válaszok megtalálására:

1. A szükséges adatok azonosítása és a források megtalálása. A potenciálisan értékes adat sok, akár nehezen hozzáférhető helyen, formában létezhet, akár a cégen belül (adatsilók az egyes céges területeken), akár a cégen kívül (közösségi média, állami adatok, fizetős adatforrások). Vizualizációs eszközökkel meg lehet vizsgálni, hogy mely adatok lehetnek relevánsak a prediktív analitikai modell megalkotásához.
2. Az adatok előkészítése a prediktív analitika egyik kihívása. A felhasználók gyakran ezzel töltik a projekt háromnegyedét: ki kell számolni az aggegált mezőket, formázni kell az adatokat, törölni a felesleges karaktereket, kezelni a hiányzó adatokat, összekapcsolni több adatforrást.
3. A prediktív modell megépítése. A megfelelő szoftver eszközökkel és a terület ismeretével be lehet vetni a szükséges statisztikai módszereket és gépi tanuló algoritmusokat a modell megalkotásához. A legjobb modell függ az adatok típusától, rendelkezésre állásától, az előrejelzés céljától. Az elemzők az adathalmaz egyik része a „tréningadatok" segítségével építik fel a modellt, majd a „tesztadatokon" értékelik a teljesítményét.
4. A modell hatásosságának és pontosságának értékelése. A prediktív analitika a valószínűségekről szól. A modell prediktív erejének a kiértékeléséhez az elemzők azt vizsgálják, hogy mennyire képes előre jelezni a tesztadatokat. Ha a modell jól működik a tesztadatokon, akkor jó jelölt a gyakorlatban történő bevezetésre.
5. A modell alkalmazásával gyakorlati lépéseket kell megfogalmazni. Az előrejelzésnek kevés értéke van, ha nem engedi a lehetőségek megragadását vagy a negatív esemény elhárítását. Az üzleti terület munkatársainak meg kell tanulniuk megbízni az előrejelzésekben, a modelleket építő adattudósoknak pedig tanulniuk kell tőlük az üzleti folyamatokról és arról, hogy milyen tudásra lenne szükség az adott folyamat vagy tevékenység javításához.
6. A modell működését, előrejelző erejét folyamatosan követni és javítani kell. A prediktív modellek annyira pontosak, mint a betáplált adatok, és idővel romlik a működésük. Az új adatokat tehát újra be kell tölteni, és problémák esetén az adatszakértőknek változtatniuk kell a modell paraméterein, esetleg új változókat kell bevonni, stb.

2.3. AZ EMC ADATELEMZÉSI ÉLETCIKLUSA

Az 5. ábra az EMC informatikai óriásvállalat adatelemzési életciklusát mutatja. Ez is hat lépésből épül fel, és sok hasonlóság figyelhető meg a Forrester előzőekben bemutatott modelljével.

5. ábra Az EMC adatelemzési életciklusa



Forrás: EMC 2013.

A folyamat során sokszor egy fázisból vissza kell lépni egy előző fázisba, mert olyan információk merülnek fel, hogy módosítani, finomítani kell az előző fázisok munkáját a felmerült új információk fényében. Ezt fejezik ki az ábrán látható ciklusok: a fázisok többször ismétlődhetnek, amíg elérik a továbblépéshez szükséges információs szintet.

1. Az első fázis a felfedezés. Ennek során kell megismerni az adott üzleti területet, illetve az előzetes eseményeket, hogy volt-e hasonló projekt a múltban, amelyből tanulni lehet az adott projekt vonatkozásában. Meg kell becsülni a projekt elvégzéséhez szükséges erőforrásokat az emberek, a technológia, az idő és az adatok vonatkozásában. Az üzleti problémát elemzési kihívásként kell átfogalmazni, amelyet aztán a következő fázisokban kell megoldani. Itt kell megfogalmazni a kiinduló hipotéziseket is.
2. A második fázis az adatok előkészítése. Ki kell alakítani az analitikai környezetet, amelyben a munka folyik a projekt során. Itt kerül sor az extrakciós, transzformációs és adatbetöltési feladatok elvégzésére, amelyek révén bekerül a rendszerbe az adat, és megindulhat az elemzési munka. Meg kell vizsgálni az adatokat, hogy aztán tovább lehessen lépni a harmadik fázisra.
3. A harmadik fázis a modelltervezés. Itt kell meghatározni a módszereket, technikákat és folyamatokat az elemzési modellek értékeléséhez. Az adatok felfedezésével, a változók közötti kapcsolatok felderítésével lehet kiválasztani a fontos változókat és a használandó modelleket.

4. A negyedik fázis a modellépítés. Itt kell kialakítani az adathalmazokat a teszteléshez, a tanításhoz és az éles alkalmazáshoz. A modellek és folyamatok futtatásához, teszteléséhez megfelelő hardver és szoftver környezetre van szükség.
5. Az ötödik fázis az eredmények kommunikációja. A felfedező fázisban megfogalmazott kritériumok alapján meg kell határozni, hogy a projekt elérte-e a kívánt célt. Meg kell fogalmazni a legfontosabb felismeréseket, számszerűsíteni kell az üzletre gyakorolt hatást, az értékteremtést, és elő kell készíteni az eredmények kommunikációját az érintett felek irányába.
6. A hatodik fázis az eredmények bevezetése. Be kell számolni a projekt eredményeiről, illetve át kell adni a kódokat és a dokumentumokat a megfelelő üzleti területeknek. Egy pilot projekt keretében be kell vezetni az eredményeket az éles üzleti rendszerekben is. Nagyon fontos az eredmények hallgatóságra szabott megfogalmazása, amely világosan kifejezi az elérhető többletértéket, és elkötelezi őket a bevezetés mellett. Ha egy technikailag pontos elemzés eredményeit nem sikerül megfelelően átadni a hallgatóságnak, akkor nem fogják látni a benne rejlő értéket, és az elemzési munka nem éri el kellő eredményt.

3. ÖSSZEHASONLÍTÁS ÉS KÖVETKEZTETÉSEK

Az egyes szerzők és szervezetek által ajánlott folyamatok között sok közös vonás figyelhető meg. Mindegyik esetben az üzleti probléma megismerésével, körüljárásával és megfelelő kérdések vagy hipotézisek megfogalmazásával indul az elemzés.

Az EMC (nagy)vállalati szemléletű tanulmányában hangsúlyos a gazdaságosság és a megtérülés szempontja is, ezért a „felfedezési" szakaszban az elemzési projekthez szükséges erőforrások megbecsülését ajánlja. Számba kell venni a szükséges eszközöket és technológiákat, illetve meg kell vizsgálni a projektben rendelkezésre álló adattípusokat. Ennek alapján lehet meghatározni, hogy elegendő-e a rendelkezésre álló adat a projektcélok eléréséhez vagy adatokat kell gyűjteni, vásárolni vagy átalakítani.

Az emberi erőforrások kiemelten fontosak. A projektcsapat összeállításánál biztosítani kell, hogy a benne szereplő üzleti szakértők, ügyfelek, elemzők és projektmenedzserek hatékony csapatot alkossanak, és meg kell becsülni, hogy várhatóan mennyi idejüket fogja lekötni a projekt. Meg kell nézni azt is, hogy az üzleti felhasználók várhatóan mennyire lesznek képesek a gyakorlatban használni a projekt eredményét, rendelkeznek-e megfelelő ismeretekkel, tapasztalattal. Ez befolyásolja az elemzési technikák kiválasztását és a bevezetés módját, formáját.

Ha nem áll rendelkezésre elegendő erőforrás a projekt sikeréhez, akkor további erőforrásokat kell bevonni. Jobb a projekt indításánál tárgyalni az erőforrásigényről, mert itt még van lehetőség a célok és megvalósíthatóság szem előtt tartása mellett biztosítani, hogy elegendő idő és erőforrás álljon rendelkezésre a megfelelő projektmunkához.

Mindegyik folyamatleírásban lényeges szakasz az adatok előkészítése, tisztítása. Általában ez a leginkább iteratív és leghosszabb időt igénybe vevő fázis. Itt történik a megfelelő informatikai környezet kialakítása, ahol meg lehet vizsgálni az adatokat a produktív rendszerek zavarása nélkül. Ha például egy projektben egy cég pénzügyi adataival kell dolgozni, akkor nem lehet a szervezet fő adatbázisának produktív verzióját használni, mert nem szabad zavarni a mindennapi riportolási tevékenységeket. Az adatokat egy kísérleti elemzési környezetben (analytical sandbox) kell összegyűjteni. Az analitikai

környezet kapacitását kellően nagyra, kb. az adott vállalat adattárházának a tízszeresére kell méretezni. Az adatforrások felé megfelelő hálózati kapcsolatokat és nagy sávszélességet kell biztosítani, hogy gyors legyen az adatok kivonása, transzformációja és betöltése.

A Big Data folyamatoknál a hagyományos extract-transform-load (ETL) műveleti sorrend helyett az extract-load-transform (ELT) sorrend a jellemző. Az ETL azt jelenti, hogy az adatok transzformációja megelőzi a betöltést, de a Big Data analitikai projektek esetében célszerűbb a betöltést a transzformáció előtt elvégezni. Ez azt jelenti, hogy az adatokat nyers formában töltik be az adatbázisba, ahol az elemzők később is eldönthetik, hogy milyen formába alakítsák át az adatokat.

A tényleges elemzés a teljes folyamat időigényének csak kisebb részét teszi ki, itt az egyes folyamatajánlások eltérnek, de lényegében analitikai modellek tervezése, futtatása, kiértékelése tartozik ide. Az adatok szerkezete és tulajdonságai befolyásolják az alkalmazandó eszközöket és technikákat (a szöveges vagy a tranzakciós adatok elemzése különböző eszközöket és megközelítéseket követel meg, mint a piaci kereslet előrejelzése pénzügyi adatok alapján).

Az EMC itt is erősen figyelembe veszi az üzleti környezet sajátosságait, is kiemeli, hogy meg kell bizonyosodni arról, hogy az elemzési technikák lehetővé teszik az üzleti célkitűzések elérését. Ehhez át kell gondolni, hogy általában hogyan oldják meg az adott problémát, hogyan keresik a választ az adott kérdésre. Meg kell nézni, hogy egy hasonló megközelítés működhet-e a rendelkezésre álló adatokkal, vagy másik megközelítésre van szükség. Sokszor jó ötleteket lehet gyűjteni más iparágban felmerült hasonló problémák megoldásából.

Mindhárom ajánlásban is megtalálható, hogy a legfontosabb változók megragadására kell törekedni, nem szabad túl sok változót bevonni a modellbe. Az egyes modellváltozatok kiértékelésével lehet azonosítani az elemzés szempontjából fontos változókat. A tesztek után a legnagyobb hatást mutató változókra kell fókuszálni, csökkentve a probléma dimenzionalitását[15].

Az EMC módszertana kihangsúlyozza az elemzési projekt eredményeinek kommunikációját, míg ez a másik két helyen alig szerepel. A projektcsapat tagjai és az érintett szereplők felé történő kommunikációban világosan ki kell emelni az elért eredményeket, de nem szabad elfeledkezni a modellhez kapcsolódó feltevésekről, a modell esetleges korlátairól, hibáiról sem.

Az utolsó szakasz a bevezetés, illetve a továbbfejlesztés. Ha az üzleti elemzés megvalósítja a kitűzött üzleti célt, és az üzleti érték, előny kellően meggyőző, akkor egy pilot projekt keretében ellenőrzött módon történhet az új módszerek, folyamatok bevezetése, majd ezt követheti a felhasználók széles köre felé történő bevezetés. Ez csökkenti a kockázatokat, mert a pilot projekt eredményeiből, esetleges problémáiból sokat lehet tanulni, majd el lehet végezni a szükséges módosításokat, mielőtt a teljes vállalatnál bevezetésre kerül az új eszköz és folyamat.

[15] Regressziós modell esetén meg kell keresni azokat a független változókat, amelyek leginkább magyarázzák az eredményváltozó változásait, és szoros kapcsolatban állnak az eredménnyel, de egymással kevéssé korrelálnak. Figyelembe kell venni a különböző adatmodellezési problémákat, mint például a kollinearitást.

Ahogy egyre többen használnak valós idejű információt és elemzéseket a napi munkában, úgy egyre nagyobb a szoftverhiba kockázata és költsége. A nagyvállalatok ezért egyre inkább törekednek az analitikai rendszerek folyamatos tesztelésére. Az üzleti folyamatok automatikus validációja felgyorsíthatja az új szoftverelemek bevezetését, és garantálhatja, hogy az éles üzleti folyamatokat nem fogja megzavarni az új technológia bevezetése. Egy komolyabb Big Data projekt esetében egy nagyvállalat akár napi több ezer döntést is alapozhat az újonnan kifejlesztett elemzési eszközre, ezért kulcsfontosságú annak biztosítása, hogy a riportok és elemzések helyesek és pontosak legyenek. Egy bizonyos méret felett ennek tesztelése kézzel már nem is lehetséges (Ballou és Marden 2014).

A Big Data technológiában rejlő üzleti potenciál realizálásához szükség van a vállalati gyakorlatban az elemzési folyamat módszeres kialakítására, mert a korábbi elemzésnél bonyolultabb lépéseket kell összehangolni a megfogalmazott célok eléréséhez. Az EMC ajánlásában nagyon jól megjelenik a vállalati környezetet, az üzleti megfontolásokat figyelembe vevő folyamatorientált szemlélet. Ez nagymértékben elősegíti a sikeres Big Data projektek végrehajtását. Sok hasonló projekt megvalósulása pedig hozzájárul ahhoz, hogy a Big Data jelenség valóban forradalmi változásokat hozzon az üzleti életben. A technológia már rendelkezésre áll, és a tanácsadó cégek, informatikai vállalatok által megfogalmazott folyamatajánlásokkal a menedzsment módszerek is nagyot léptek előre.

4. ÖSSZEFOGLALÓ

A Big Data elemzési feladatok száma és jelentősége gyorsan növekszik napjainkban. A nagymennyiségű és különböző adatok kezelése számos nehézséggel jár, és az elemzési folyamat lényegesen komplexebb, nehezebb és hosszabb, mint a korábbi, „hagyományos” elemzési feladatok esetén.

Egy jól definiált elemzési folyamat segíti a komplex probléma lebontását kezelhető részekre, felhívja a figyelmet lényeges, de könnyen elmaradó feladatokra, és ezáltal nagyban hozzájárul a projekt eredményességéhez. Több informatikai cég is felismerte ennek a jelentőségét, és ajánlásokat dolgoztak ki a Big Data elemzési folyamatokhoz. Jelen cikkben a Fayyad és szerzőtársai klasszikus cikkében, valamint a Forrester Research és az EMC tanulmányában megfogalmazott ajánlásokat tekintettük át és hasonlítottuk össze.

A projektfolyamat felépítésére adott javaslatok között sok a hasonlóság. A legfontosabb közös elem az, hogy a tényleges elemzés és informatikai fejlesztés előtt alaposan meg kell ismerni az üzleti problémát, meg kell fogalmazni a lényeges kérdéseket, az elérendő üzleti célokat, hogy aztán az elemzési projekt valóban értéket teremtsen a vállalat számára. A tényleges elemzési munka csak kisebb részét teszi ki a projekt idő- és erőforrásigényének, az adatok megszerzése és előkészítése jellemzően sokkal nagyobb feladat.

A projekt eredményeinek gyakorlati hasznosulását nagymértékben befolyásolja a megfelelő kommunikáció a döntéshozók felé, illetve a bevezetés menete. A bevezetéshez kapcsolódó kockázatok lényegesen csökkenthetők egy pilot szakasszal és megfelelő teszteléssel. A fejlesztések számának és bonyolultságának növekedésével egyre inkább szükség van a tesztelési folyamatok automatizációjára. Egy nagyvállalat akár napi több ezer döntést is alapozhat az újonnan kifejlesztett elemzési eszközre, ezért kulcsfontosságú

annak biztosítása, hogy a riportok és elemzések helyesek és pontosak legyenek. Egy bizonyos méret felett ennek tesztelése kézzel már nem igazán lehetséges.

5. IRODALOMJEGYZÉK

Ballou, M. C. – Marden, M. (2014): The Business Value of Worksoft Automated Business Process Validation Solutions, *IDC white paper*, August, 2014. Framingham, MA, USA.

Codd, E. F., Codd, S. B. & Salley, C. T. (1993): Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate E. F. Codd and Associates .

EMC (2013): Data Science and Big Data Analytics Student Guide, *EMC Education Services publication*, June 2013. Hopkinton, MA, USA.

Fayyad, U. – Piatetsky-Shapiro, G. – Smyth, P. (1996): From data mining to knowledge discovery: an overview. *Advances in knowledge discovery and data mining*, pp. 1-34. American Association for Artificial Intelligence. Menlo Park, CA, USA. http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf, letöltés: 2015.06.15.

Felden, C. (2015): Business Analytics előadás, jegyzetek. *TU Bergakademie Freiberg*, Freiberg (Sachsen)

Gualtieri, M. – Curran, R. (2015): The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015, *Forrester Research*, April 1st, 2015. https://www.predixionsoftware.com/Portals/0/Analyst%20Reports/The%20Forrester%20Wave_%20Big%20Data%20Predictive%20Analytics%20Solutions_%20Q2%202015.pdf, letöltés: 2015.05.18.