DOI https://doi.org/10.15407/csc.2023.03.033
УДК 004.65:004.75:004.738.5

О.А. УРСАТЬЄВ, к.т.н., с.н.с., провідний науковий співробітник, Відділ комплексних досліджень інформаційних технологій, Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
ORCID: https://orcid.org/0009-0009-8323-0525,
03187, м. Київ, просп. Академіка Глушкова, 40, Київ, Україна,
aleksei@irtc.org.ua

ДОСЛІДЖЕННЯ ДАНИХ У ПРОМИСЛОВИХ DATA MINING ПРОЄКТАХ В ЕПОХУ ГЕНЕРАЦІЇ ВЕЛИКИХ ДАНИХ

Розглянуто еволюційні зміни платформ виявлення необхідної інформації в даних та їхнє подальше аналітичне дослідження в промисловості та інших сферах діяльності суспільства. У зв'язку з неухильним збільшенням усіх доступних типів даних традиційний бізнес-аналіз із залученням IT-служб не задовольняє компанії. Бізнес потребує значно меншого часу для розуміння, щоб залишатися конкурентоспроможним і знаходити нові можливості свого розвитку. Наведено підходи, аналітичний апарат та інструментарій для отримання інформації, орієнтовано на бізнес-користувачів.

Ключові слова: *рівні зрілості аналітики, інтелектуальна аналітика, виявлення інформації в даних (Data Discovery), дослідження даних (Data Science), підготовка даних (Data Preparation), дані самообслуговування (Self Service Data), платформи підготовки даних на засадах самообслуговування, конвеєр підготовки даних.*

Вступ

Неухильне зростання обсягів вироблених спільнотою (промисловість, економіка, бізнес тощо) Великих Даних (*Big Data*) [1], їхня різноманітність і складність внаслідок нетрадиційних форматів, змушують у певних ситуаціях використовувати лише ті з них, які вкрай необхідні для подолання окремих проблем бізнесу, замість того, щоб докладати величезних зусиль для інтеграції значних обсягів різних даних. Комбінування типів даних (*mashup data*) є складним і працемістким технічним завданням. Водночас використання великих даних суттєво підвищує ефективність функціонування та конкурентоспроможність організацій приватного й державного секторів, керування соціальними проєктами та дає змогу створювати приріст національних і світових економік. На сучасному етапі дані перетворюються на капітал, який безпосередньо бере участь у формуванні та керуванні різними сферами діяльності людини.

Ефект від використання великих даних із різноманітних джерел суттєво відрізняється. Зазвичай його сховано в нетрадиційних даних. Завдання полягає у визначенні того, що є вартісним, а потім відбувається добування та трансформування цих даних для аналізу ра-

зом із наявними історичними даними, тому що такий комбінований набір підвищить цінність внаслідок синергетичного ефекту (греч. *synergos* – що діє сукупно). Кінцеву значущість великих даних буде оцінено з урахуванням корисності та точності інформації і зменшенням часу одержання відповіді [1].

Суть цінності даних також полягає в їхньому необмеженому повторному використанні – альтернативній цінності. Абсолютна цінність даних може набагато перевищувати ту, яку вдається видобути під час первинного використання. Інноваційні компанії можуть отримати приховану цінність і потенційно отримати величезні переваги, тобто, цінність даних треба розглядати як можливості їхнього подальшого використання, а не лише нинішнього. Яскравим прикладом вилучення прихованої цінності з використаних даних слугує матеріал [2] Інтернет-компанії *Google* в науковому журналі *Nature*, де оприлюднено один зі способів раннього виявлення епідемії грипу за допомогою моніторингу звернень за медичною допомогою у формі онлайн-запитів до пошукових систем. Оскільки відносна частота певних запитів відчутно корелює з відсотком відвідувань лікаря, коли пацієнт має грипоподібні симптоми, можна оцінити поточний рівень тижневої активності грипу в кожному регіоні з відставанням приблизно на день. Прогноз компанії *Google* ґрунтується на повторному аналізі великого набору даних, використаних раніше і збережених. Інший приклад вилучення прихованої цінності – корисна інформація з цифрового сліду («викиду даних»), побічного продукту інших видів діяльності спільноти користувачів [1]. На основі зібраних даних поліпшують наявні або розробляють нові служби із застосуванням принципу рекурсивного навчання. Кожну дію користувача вважають сигналом, який аналізують і повертають до системи навчання. Викиди даних – це механізм, покладений в основу багатьох комп'ютеризованих служб, таких як розпізнавання голосу, спам-фільтри, перекладачі тощо. Коли користувач зазначає в програмі розпізнавання голосу, що вона неправильно зрозуміла сказане слово, він, по суті, навчає систему, вдосконалюючи її.

Тема: роботу присвячено огляду інформаційних послуг, що їх надають служби *IT* компаніям, підприємствам, тощо з ведення бізнесу загалом. Технології надання послуг, природно, змінюються з часом і під впливом конкуренції в бізнесі, і внаслідок розвитку інтелектуальних засобів обробки інформації.

Мета: ознайомити фахівців, що потребують навичок роботи з даними, із завданнями, пов'язаними з обробкою даних, досконалим математичним апаратом та інструментарієм світового рівня для отримання з них інформації та знань.

Business Intelligence (*BI*). Традиційні *BI* та рівні зрілості аналітики

Виявлення «потрібних» даних, що зумовлюють найважливіші неінтуїтивні бізнес-ідеї з усіх доступних користувачеві типів даних, з подальшим аналізом впливу значущої сукупності даних на бізнес-об'єкт, раніше виконували *BI* із залученням *Data Mining* [3]. Як відомо, словосполука *Business Intelligence* є неоднозначною [3]. Перша поява абревіатури *BI* датується 1958 р. [4], і в перекладі з англійської вона може означати швидке розуміння того, що відбувається, бізнес-аналіз, аналітику і навіть інтелект. Проте слово *intelligence* означає вміння міркувати розумно, а не інтелект, для якого є термін *intellect*. Тож у понятті *BI* слово *intelligence* розуміють як «здатність сприймати взаємозв'язок представлених фактів у такий спосіб, щоб спрямовувати знання на досягнення бажаної мети». Бізнес представлено в широкому сенсі колекцією заходів, що їх здійснюють із будь-якою метою *BI* науці, технології, торгівлі, промисловості тощо. Лише з 90-х років XX ст. термін *BI* почали обговорювати внаслідок повторного відкриття для ремаркетингу систем *DSS* [3], що їх застосовували з 60-х до середини 80-х років. Вважають, що *BI* еволюціонував від систем *DSS*, створених для сприяння прийняттю рішень та плануванню.

Термін став відомий внаслідок впливу *Gartner*[1] [5–7]. Авторство реанімованого терміна *BI* приписують Говарду Дреснеру[2], який запропонував його як «парасольковий» термін для опису «концепцій і методів для покращення прийняття бізнес-рішень із застосуванням систем підтримки, ґрунтованих на фактах». Пізніше Говард Дреснер уточнив визначення *BI* як «загальне поняття, що описує концепції та методи збирання, інтегрування, аналізування та подання даних, необхідних для прийняття обґрунтованих управлінських рішень за допомогою *IT*-підрозділу» [8].

Спочатку *BI* було орієнтовано на збирання інформації та підготовлення регламентованої звітності, а їхні аналітичні можливості було обмежено здатністю надавати традиційні функції запитів. Багато структур даних, які з'явилися в 1980-х роках у стилі *OLAP*, що містять *MOLAP, ROLAP* і *Hybrid-ROLAP* [3], застосовували *OLAP*-інструменти для всебічного аналізування даних у *BI* (*on-line analytical processing* – оперативне аналітичне оброблення) [3], які допускали розгляд різних зрізів даних, зокрема тимчасові, що дають змогу виявляти різні тренди та залежності (за регіонами, продуктами, клієнтами тощо). Багатовимірний аналіз даних на базі технології *OLAP*-кубів дав змогу самостійно, не покладаючись на *IT*-службу, створювати нерегламентовану звітність на основі власних запитів до даних. Підтримка *OLAP*-кубів значно пришвидшила оброблення запитів і дала змогу користувачам аналізувати дані, забезпечуючи стилі аналізування, відомі як «нарізка на скибочки та кубики – *slicing and dicing*». Це означало спосіб сегментування в поздовжньому та поперечному напрямку для перегляду та розуміння даних. Користувачі почали переміщатися в багатовимірному (*multidimensional drill paths*) просторі, видобуваючи дані. У них з'явилася можливість зворотного запису значень у базу даних для планування та моделювання простими бізнес-правилами ситуації (*if-then*)[3] «якщо – то» тощо.

Традиційні *BI* характеризують два ключові моменти: звична функція запитів (*SQL*-запити), онлайн-аналітичне оброблення (*OLAP*) та спеціалізовані запити. Це дало змогу користувачам отримувати довільну або спеціальну звітність (*ad-hoc reporting*), яка допомогла одержувати за необхідності додаткову інформацію та розгорнуту звітність (*drill-down reporting*) з послідовним збільшенням рівня детальності даних, що їх розглядають, самостійно, не залучаючи фахівців *IT*-служб та аналітиків, заглибитися в аналізований процес, розібратися та зрозуміти зміст представлених і використовуваних да-

[1] *Gartner, Inc* провідна світова дослідницька і консалтингова компанія з кількістю працівників 19 500 (грудень 2022 р.), розташованих у більш ніж 100 офісах світу. Була заснована в 1979 році Гідеоном Гартнером (*Gideon Gartner*) як постачальник замовних та тиражних досліджень і аналізу в галузі інформаційних технологій (*IT*) зі спеціалізацією на інформації про *IBM* та її продукти. У той час Гартнер уже мав багаторічний досвід роботи в галузі *IT*, після того як він працював у *IBM* як у Сполучених Штатах, так і в Європі у відділі конкурентної розвідки компанії. Має ступінь бакалавра з машинобудування (Массачусетський технологічний інститут) та ступінь магістра з менеджменту. Відомим надбанням компанії є розроблене *Gartner "Magic Quadrant, MQ"* – Магічний квадрант, тип звіту про дослідження ринку, та "*Hype cycle*" – "Цикл очікувань" або "Цикл *Gartner*" – крива входження або зрілості інноваційних технологій. *MQ* – це серія звітів про дослідження ринку, опублікованих *IT*-консалтинговою компанією *Gartner*, які покладаються на власні методи аналізу якісних даних для демонстрації ринкових тенденцій, таких як напрямок, зрілість і учасники. Їхні аналізи проводяться для кількох конкретних технологічних галузей і оновлюються кожні один–два роки. Цикл очікувань *Gartner* – це графічна презентація, розроблена, використана та брендована американською дослідницькою, консультативною та інформаційною компанією *Gartner*, щоб показати зрілість, впровадження та соціальне застосування конкретних технологій. Цикл очікувань стверджує, що забезпечує графічне та концептуальне представлення зрілості нових технологій у п'ять етапів. Модель піддавалася критиці з різних причин, у тому числі через ненаукову точність і використання суб'єктивної термінології. *Gartner Wikipedia* https://en.wikipedia.org/wiki/Gartner; *Gartner hype cycle* https://en.wikipedia.org/wiki/Gartner_hype_cycle.

[2] *Говард Дреснер* (*Howard Dresner*) відомий тим, що у 1989 р. використовував термін «*Business Intelligence*», працював 13 років у *Gartner* як науковий співробітник та провідний аналітик.

[3] *SAS Blogs. Why analytics is better than simple if-then business rules, Febr.* 27, 2014 https://blogs.sas.com/content/sascom/2014/02/27/why-analytics-is-better-than-simple-if-then-business-rules/

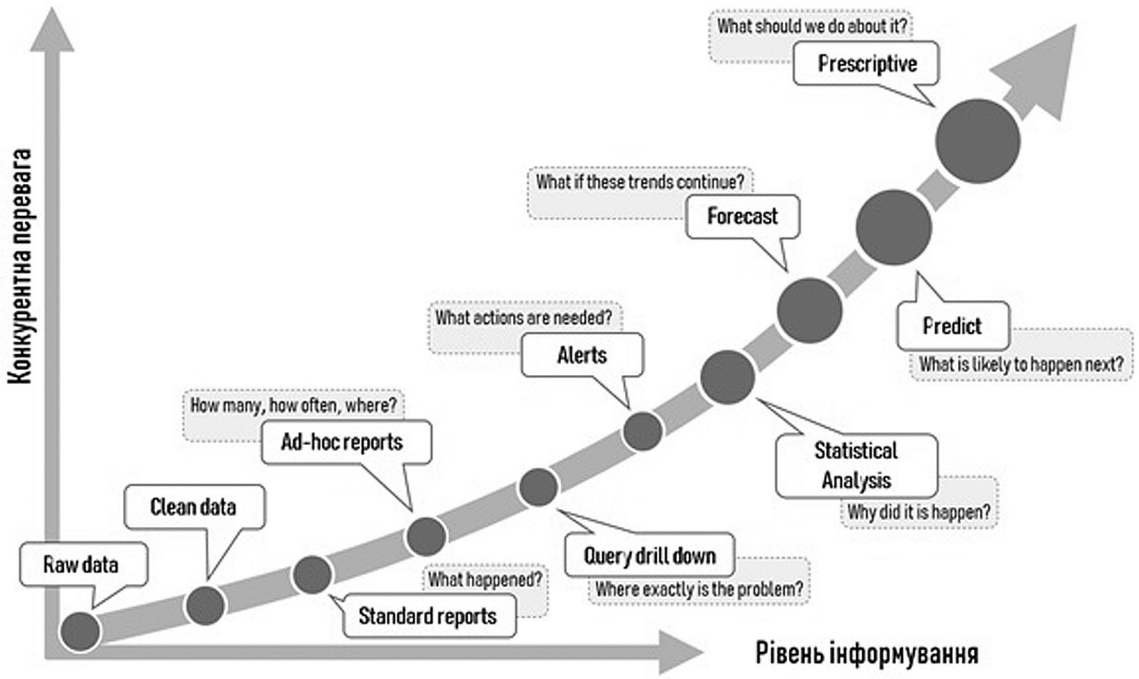

Рис. 1. Рівні зрілості аналітики

них. Ідея *OLAP* завжди передбачала мету бізнес-аналітики «самообслуговування», в якому бізнес-користувачі отримують доступ до даних без підтримки *IT*.

На цьому етапі розвитку *BI* застосовується описова аналітика (*Descriptive analytics: What happened*?), що відповідає на питання: «Що сталося?» (рис.1). Переважна більшість застосунків на платформах *BI* надавали лише узагальнювальні статистики (матсподівання, дисперсія, медіана тощо). Пізніше до них долучилися кластеризація та самоорганізовані карти (*self-organizing maps – SOM*)[4], афінність (аналіз спорідненості) та аналіз графа (*graph analysis*), спільний та оглядовий аналізи; оцінювання щільності; метрики подібності, а також оброблення неструктурованих даних та текстова аналітика. Описову аналітику часто асоціюють з візуалізацією даних через звіти, інформаційні (контрольні) панелі та системи показників. Інтерактивна діагностична візуалізація дала змогу ефективніше відображати численні аспекти даних, застосовуючи зображення та діаграми замість звичних таблиць – максимально наочне, зрозуміле подання даних із застосуванням різних шкал, показників, зображень, схем та графіків. Інтуїтивно зрозумілі візуалізації та інтерфейс користувача сприяли повсюдному впровадженню аналітики. У сукупності все це є невід'ємні складники передових технологій *BI* того часу.

Еволюційні зміни досліджень даних

Свій *Magic Quadrant* (*MQ*) для *BI* та аналітики *Gartner* розробляє з 2006 р., проте досі завжди йшлося про спроможності програмних засобів цього класу у сфері традиційних запитів і звітів. Описову аналітику переважно було завершено для більшості великих компаній у традиційних сферах, таких як фінанси та продажі, проте компанії середнього бізнесу здебільшого на початок 2013 р. ще тільки розпочали впровадження *BI* систем [9–15].

[4] Популярна нейромережева архітектура, орієнтована на навчання без учителя. Кохонен Т. Самоорганизующиеся карты. М.: БИНОМ. Лаборатория знаний, 2008. 655 с.: (Адаптивные и интеллектуальные системы) – https://docplayer.ru/50012410-Samoorganizuyushchiesya-karty.html

Тим часом, протягом кількох минулих років інтерактивна діагностична візуалізація, як і прогнозний та рекомендаційний аналізи, ставали дедалі зрозумілішими та важливішими для організацій. Домінантною темою світового ринку *BI* у 2012 р. стала аналітична архітектура з так званою розширеною або поглибленою аналітикою (*Advanced Analytics*), до якої відбувся багаторічний перехід ринку *BI* й аналітичних платформ від *IT*-звітності до сучасних *BI*, орієнтованих на бізнес-користувачів. За *Gartner* [10, 11] – це «аналіз усіх видів даних із застосуванням складних кількісних методів (статистика, описовий та інтелектуальний прогнозний і рекомендаційний аналізи даних, моделювання й оптимізація тощо) для одержання інформації про те, що традиційні підходи до бізнес-аналітики (*BI*) – такі як запити та звітність – навряд чи виявлять» [11]. Там же висловлено думку, що «інтелектуальна аналітика та інші категорії просунутої аналітики стають головним чинником на ринку аналітики». Як зазначено у звіті *Gartner «Magic Quadrant for Business Intelligence and analytics platforms*» за 2013 р. [9], пропозиції *Advanced Analytics* охоплюють різні варіанти застосування системи *BI* та рівні зрілості аналітики (див. рис. 1), що містять фази: інтерактивна діагностична візуалізація (*Interactive Diagnostic Visualization*), прогностична або предиктивна (*Predictive Analytics*) і рекомендаційна (*Prescriptive Analytics*) аналітики.

На відміну від базової описової аналітики розширена аналітика забезпечує детальний аналіз інформації, відповідаючи на запитання «Чому це сталося? / *Why did it is happen*?» і «Що нам із цим робити? / *What should we do about it*?». Можливість ухвалити рішення, підкріплене обґрунтованим аналізом даних, здатна суттєво покращити бізнес-результати. Наприклад, клінічна діагностика застосовує аналітику серед широкого діапазону структурованих та неструктурованих джерел, зокрема електронні записи про стан здоров'я, інформацію про раніше вжиті ліки, діагностичні та інші дані. Усе це дає змогу призначити оптимальне лікування кожного пацієнта. Але для отримання реально цінних відомостей із даних потрібні знання: від інтеграції та підготовки інформації до глибинного аналізу, спеціалізованих обчислювальних середовищ та інтелектуальних алгоритмів. Щоб досягти прийнятних для бізнесу результатів, потрібні фахівці, які мають відповідні компетенції та досвід [9, 16].

Predictive analytics – це клас методів аналізування даних, що дають змогу прогнозувати деякі невідомі події чи майбутні результати, спираючись на відомі дані. Структури, які можна знайти у відомих даних (територія *descriptive analytics*), фактично є додатковими даними, спочатку не відомими, але які можна отримати з відомих даних за допомогою прогнозних моделей (класифікації, кластеризації, часових рядів, регресії тощо). Мета полягає в тому, щоб передбачити невідому інформацію за відомими даними. Прогнози екстраполюють з історичних даних, і вони залежать від якості та різноманітності відомих даних [17].

Основне завдання прогностичної аналітики – визначення одного чи кількох чинників (прогностичний фактор, предиктор), які впливають на передбачувану подію. Прогнозні моделі застосовують шаблони (*pattern*, патерни), отримані в історичних та актуальних даних для оцінювання ризику або можливості прогнозування, пов'язаного з певним набором умов. Імовірність настання певної події в майбутньому або отримання бажаного результату оцінюють стосовно кожного окремого чинника моделі й ця імовірність слугує для інформування або впливу на організаційні та виробничі процеси (наприклад, виявлення шахрайства, охорона здоров'я, виробництво тощо) [18, 19].

Forecasting (прогнозування) – це конкретний тип прогнозування (див. рис.1) із застосуванням аналізу часових рядів або економетричних методів, на основі яких можна будувати моделі ймовірнісного розвитку подій «Що, якщо ці тенденції продовжаться? / *What if these trends continue*?» та давати прогноз результатів цього розвитку, наприклад, для прогнозування

значення змінної або результату в зазначений час, наприклад, продажу в наступному кварталі або кількості дзвінків, які колл-центр отримає наступного тижня [15].

Simulation (імітаційне моделювання) – це підхід прогностичної аналітики, який охоплює побудову моделі об'єкта (процесу) для експериментування чи вивчення його функціювання, з наголосом на розуміння діапазону можливих результатів.

Опис процесу прогнозного аналізу наведено в [19, 20]. Там, як приклад наведено деякі прогностичні моделі та розглянуто зарубіжні, широко розповсюджені прогресивні аналітичні засоби, а також застосування в них одного з найефективніших методів індуктивного моделювання та прогнозування – вітчизняного методу групового врахування аргументів (МГУА), сучасного програмного засобу прогнозної аналітики на базі класичного алгоритму МГУА, що виконує прогнозування часових рядів, розв'язання задач класифікації та кластеризації.

Аналітика класу *advanced* застосовує статистику, описові та предиктивні інструменти інтелектуального аналізу даних і тексту, моделювання, зокрема й фінансове, економетрику та оптимізаційні засоби, наприклад, математичний апарат теорії ігор у моделі оптимізації. Кінцевою метою застосування інструментів удосконаленої аналітики є прийняття рішень у бізнес-завданнях, ідентифікація можливостей для складання адекватних прогнозів, виявлення нових процесів, шаблонів та інших закономірностей.

Прогнозна аналітика передбачає [14] індуктивне, регресійне моделювання; аналізування часових рядів; нейронні мережі, що їх широко застосовують під час аналізування даних [18]; ситуаційне моделювання; дерева класифікації та регресії – скорочено *CART*. Цей термін введено для позначення алгоритмів дерева рішень, що їх застосовують у задачах класифікації або прогнозного регресійного моделювання; подальші техніки індукції[5] правил; базовані на екземплярах[6] підходи; Байєсівське моделювання; ансамблі[7] та ієрархічні моделі; розробка інших прогностичних моделей; тестування прогностичних моделей тощо. Такі методи як регресія, дерева рішень та ансамблеві моделі застосовують для оцінювання поведінки чи неочікуваних результатів. Це також охоплює можливість порівняння та перевірки моделей під час їх обирання [13].

Основна можливість сучасних аналітичних платформ – полегшення синтезу моделей за експериментальними даними в умовах прогностичної невизначеності та неповноти вихідної інформації. Таке завдання зводиться до побудови математичної моделі, що наближає невідому закономірність функціонування досліджуваного об'єкта (процесу), за інформацією, яка неявно міститься у вибірці наявних даних [13, 21].

Актуальні технології прогностичної аналітики ґрунтуються на статистичних методах, зокрема, методах прогнозного моделювання, машинного навчання та інтелектуального аналізу даних, які застосовують для дослідження поточних та історичних фактів, що характеризують спостережуваний процес (об'єкт), щоб передбачити, з певним ступенем довірчої ймовірності, майбутні чи інші невідомі події [18–20].

Prescriptive analytics (рекомендаційна аналітика) [9, 13, 15, 17, 22, 27] – третя, завершальна фаза бізнес-аналітики, створена на вершині описової та прогностичної аналітик (див. рис.1). Остання, маючи статистичні моделі та методи прогнозування, забезпечує відповідь на випереджальне (*forward-looking*) питання: «Що правдоподібно може статися далі? / *What is likely to happen next*?» і допомагає передбачити можливі сценарії активного періоду, а не покладається на реактивний підхід до прийняття рішень. Обробляючи історичні дані,

[5] Індукція правил – це метод, який створює правила типу "if-else-then" ("якщо-тоді") з набору вхідних і вихідних змінних.

[6] Застосовують у машинному навчанні під час моделювання складних завдань прийняття рішень.

[7] Ансамбль методів у статистиці та машинному навчанні використовує кілька навчальних алгоритмів з метою отримання кращої ефективності прогнозування.

прогностична аналітика «навчається» на минулих подіях і створює моделі, які можуть бути застосовані для прийняття рішень щодо поточних або майбутніх подій.

Розкриваючи шаблони, взаємодії та зв'язки, приховані в даних, прогнозна аналітика забезпечує розуміння ймовірних сценаріїв і майбутні результати. Поєднання прогнозної аналітики та методів оптимізації забезпечує можливість перетворити розуміння на дію.

Рекомендаційна аналітика просуває прогнозну аналітику на крок вперед, пропонуючи конкретні дієві поступи. У той час, як прогностична аналітика може передбачити, що станеться, коли й чому настане ця подія, рекомендаційна аналітика запропонує варіанти можливостей або зменшення майбутніх ризиків, а також потенційних результатів кожного рішення. По суті, цю аналітику призначено для надання рекомендацій; вона намагається кількісно оцінити вплив майбутніх рішень і дати прогноз про можливі результати їх здійснення. Отже, прогнозна аналітика надає сировину для прийняття обґрунтованих рі-шень, тоді як рекомендаційна аналітика на основі сукупності даних надає варіанти рішень, які можна порівняти між собою, обрати найкраще у відповідь на питання: «Що нам належить із цим зробити? / *What should we do about it*?».

Прогностичній аналітиці бракує аналітично виважених, оптимізованих наборів рішень. Рекомендаційну аналітику вважають найцінішою частиною аналітики, оскільки її можна застосовувати для пропонування конкретних рішень, необхідних для досягнення бажаного результату в бізнесі. Це прийняття складних рішень, обтяжених компромісами між бізнес-цілями та наявними обмеженнями, із застосуванням прогностичного моделювання, технології оптимізації та додавання евристики. За допомогою предиктивної аналітики можна спрогнозувати множину варіантів результату події або поведінки клієнта. При цьому в наявності вже буде кілька варіантів реагування, що в мінливих умовах дасть змогу швидко прийняти оптимальне рішення. Роль рекомендаційної аналітики полягає в тому, щоб обрати прогнозну модель, яка забезпечує досягнення найкращого результату, з огляду на всі відомі ресурси, минулі результати й те, що відбувається в цей визначений час. Рекомендаційне моделювання підтримує прийняття рішень, які забезпечують розуміння інформації та подій у режимі реального часу, дає змогу виокремити правильне значення прогностичного чинника (чинників) на основі набору обмежень для детермінованих процесів і за допомогою імітаційного моделювання результатів для випадкових процесів (рис. 2).

Застосовують розширену аналітику в різний спосіб. Один із них полягає в застосуванні досконаліших, функційно розвинених аналітичних платформ, що дасть змогу *IT*-фахівцям створювати практично будь-які аналітичні моделі стосовно визначених завдань. Вони надають достатній набір інструментарію для досвідчених користувачів, традиційно – фахівців з даних (*data scientists*). Однак дедалі частіше відбувається орієнтація на бізнес-аналітиків і менш кваліфікованих фахівців з даних, які, проте, добре обізнані з предметною областю (ПрО) – «*citizen data scientist*»[8]. З іншого боку, швидке збільшення обсягу доступних даних, і особливо їхніх нових різновидів (таких як неструктуровані дані про взаємодію з клієнтами та потокові дані, генеровані обчислюваною технікою та міжмашинною взаємодією – фізичне підґрунтя *IoT*, тощо), потребує швидкого інтерпретування даних і реагування на інформацію, що міститься в них. Підвищений попит на ці типи можливостей, що випереджає пропозиції досвідчених користувачів, потребує вищого рівня автоматизації й формує потребу в самообслуговуванні даних та в інструментах, зрозумілих фахівцям у ПрО – це другий спосіб реалізації аналізу різних типів даних.

[8] Експерт в ПрО у *Gartner* «*citizen data scientists*» – це дослівно цивільні спеціалісти з даних. *Gartner* визначає *Citizen Data Scientist* як «особу, яка створює або генерує моделі, що використовують прогнозну або рекомендаційну аналітику, але чия основна робоча функція перебуває поза цариною статистики та аналітики» – 20 квіт. 2018 р.

Сфери з інтенсивним використанням даних, такі як охорона здоров'я, промисловість, фінансові послуги, урядування тощо, можуть отримати вигоду із застосування рекомендаційної аналітики. Це особливо цінно через високу вартість людських помилок. Однак рекомендаційна аналітика не є, безумовно, абсолютно надійною. Вона є ефективною, якщо організації знають, які запитання ставити та як реагувати на відповіді. У тому разі, якщо вхідні припущення є неправильними, результати не будуть точними. Однак за ефективного застосування рекомендаційна аналітика може допомогти приймати рішення на основі ретельно проаналізованих фактів, а не робити поспішні висновки на основі інтуїції. Рекомендаційна аналітика може імітувати різні результати й оцінювати ймовірність кожного, допомагаючи краще зрозуміти рівень ризику та невизначеності (зрозуміти ймовірність найгірших сценаріїв) [27].

Сфери з інтенсивним використанням даних, такі як охорона здоров'я, промисловість, фінансові послуги, уряд тощо, можуть отримати вигоду із застосування рекомендаційної аналітики. Це особливо цінно через високу вартість людських помилок. Однак рекомендаційна аналітика не є, безумовно, надійною. Вона ефективна, якщо тільки організації знають, які питання ставити та як реагувати на відповіді. У разі, якщо вхідні припущення неправильні, результати не будуть точні. Однак за ефективного застосування рекомендаційна аналітика може допомогти приймати рішення на основі ретельно проаналізованих фактів, а не робити поспішні висновки на основі інтуїції. Рекомендаційна аналітика може імітувати різні результати та оцінювати ймовірність кожного, допомагаючи краще зрозуміти рівень ризику та невизначеності (зрозуміти ймовірність найгірших сценаріїв) [27].

Data Discovery

Виявлення інформації в даних (*Data Discovery*) – це ітеративний процес пошуку закономірностей і викидів за допомогою візуальної навігації за даними та застосування розширеного аналітичного апарата для отримання інформації з можливістю забезпечення легкого керованого доступу до мультиструктурованих даних, орієнтований на бізнес-користувачів. Дослідження даних (*Data Science*) потребує навичок розуміння взаємозв'язків даних і моделювання даних, а також застосування функцій аналізування даних і спрямованої розширеної аналітики для отримання інформації [28, 29]. У звітах *Gartner* під *Data Science Platform* розуміють досконаліше, під кутом зору застосовуваної аналітики, оброблення даних до проникнення в їхню суть на основі машинного навчання [30, 31] та штучного інтелекту [32].

Звернімо увагу на суперечливість понять, що стоять за терміном *Data Science* в англійській та російській: «наука про дані» мовах. *Data Science* – це міждисциплінарна область, у якій застосовують наукові методи, процеси, алгоритми та системи для отримання знань (*knowledge*) зі структурованих та неструктурованих даних та їхнє розуміння (*insights*). Інтелектуальний аналіз даних та великі дані є підрозділами. *Data Science* – це «концепція об'єднання статистики, аналізу даних, машинного навчання та пов'язаних з ними методів» для «розуміння та аналізування реальних явищ», наданих даними. Вона застосовує методи та теорії, взяті з багатьох галузей у контексті математики, статистики, комп'ютерних наук та інформатики[9]. І завершу цитатою Хела Варіана (*Hal Varian*), яка відбиває такі аспекти: «Здатність брати дані – вміти розуміти їх, обробляти, схоплювати їхню цінність, візуалізувати та передавати їх – це стане надзвичайно важливою навичкою наступними десятиліттями[10]». Отже, задача[11], розв'язувана тими, хто займається *Data Science*, полягає у добуванні знань із даних з метою

[9] *Data science* ¬– https://en.wikipedia.org/wiki/Data_science.

[10] *Loukides Mike*. 2010. What is data science? – https://www.oreilly.com/ideas/what-is-data-science.

[11] *Data science* ¬ https://www.it.ua/ru/knowledge-base/technology-innovation/data-science-nauka-o-dannyh.

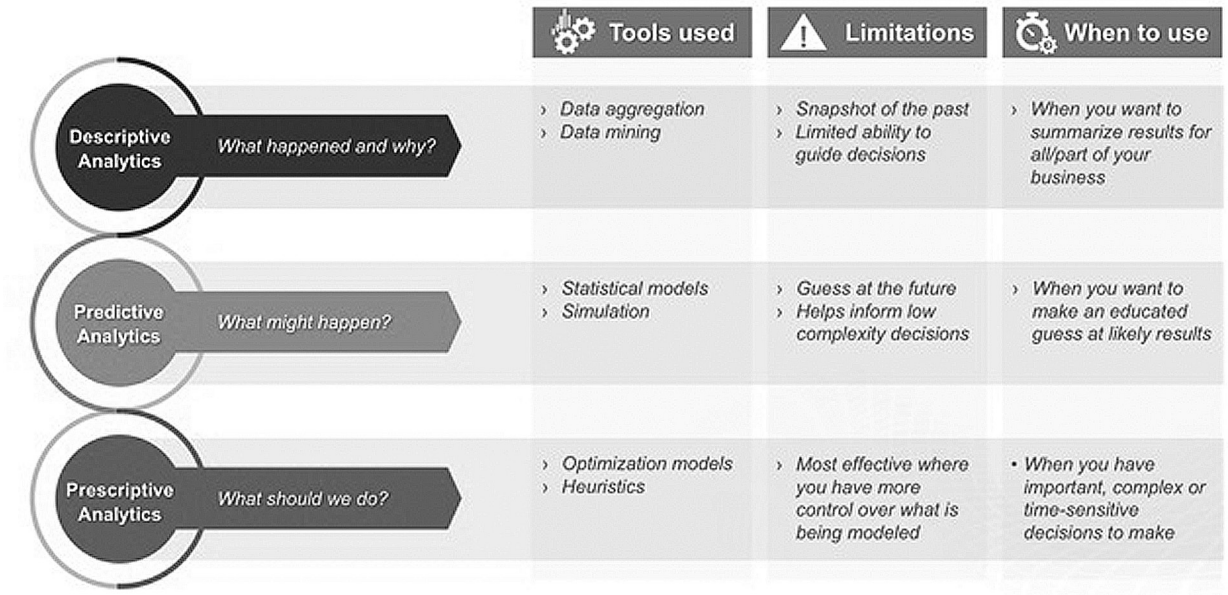

Рис. 2. Застосування аналітик

розуміння того, що містять дані, а самі дані не є для *Data Science* предметом цієї науки.

Gartner розглядає *data discovery* як альтернативу традиційному бізнес-аналізу даних із залученням *IT*-служб для отримання глибшого розуміння, що виходить за рамки звітів. Це новий напрям в області аналізу, орієнтований на бізнес-користувачів. Останнє пояснюється тим, що бізнес потребує значно коротшого, ніж будь-коли раніше, інтервалу часу для розуміння в широкому сенсі, щоб залишатися конкурентоспроможним на ринку і знаходити нові можливості свого розвитку. За традиційного аналізу, якщо користувач хоче задати додаткові запитання щодо даних, він має звернутися до фахівця з даних або з аналітики *SQL*. Але, внаслідок зайнятості цих фахівців, отримання необхідних користувачеві даних, може зайняти години або навіть дні. Посадові особи, які приймають рішення, вимагають своєчасного доступу до даних, які їм потрібні для якісного виконання своєї роботи, а також мірою того, як компанії починають усвідомлювати ціну затриманих або недостатніх даних. Тому ширше коло бізнес-користувачів потребує доступу до інтерактивних стилів аналізу та висновків розширеної аналітики, які не потребують від них навичок у галузі *IT* або оброблення даних. Відбувається перехід від застосування базових традиційних *IT*-орієнтованих платформ, які є корпоративним стандартом, до децентралізованішого розгортання на підприємстві. Реалізовані в них функції забезпечують: виявлення інформації в даних, можливвсть самообслуговування та доступ до складних, але доступних для бізнесу інструментів підготовки даних тощо.

Рішення типу *data discovery* пропонують користувачам інтерактивний графічний користувацький інтерфейс із відповідною графікою, який базується на архітектурі *in-memory*, використовують пам'ять сучасного аналітичного обчислювача як альтернативу традиційному сховищу даних із *IT*-орієнтованою системою записування (*SOR-System Of Record*) та інше, що відповідає запиту бізнесу на прості та швидкі в роботі системи аналізу. Ці рішення не потребують значної участі *IT*-фахівців у попередній підготовці даних як обов'язкової умови для аналізу. Технологія *in-memory* полегшує етап обстеження, а також забезпечує швидке його виконання.

Виявлення даних, зазвичай, відбувається в «озерах даних» – *Data Lake* – корпоративному сховищі сирих даних у вхідному форматі з різноманітних джерел або у сховищах орга-

нізацій (*silos*) на рівні підрозділів, де аналітики виконують дослідження, чи в середовищі великих даних загалом. Коли дослідження не дають позитивного ефекту, тобто не додають нового результату (*when these analytics do not add up*), компанія витрачає дорогоцінний час на додаткові джерела даних, намагаючись отримати загальну картину того, що відбувається [33].

Teradata Aster Discovery є прикладом промислової платформи виявлення інформації в даних [34], яка підтримує БД T*eradata Aster* з портфоліо *Discovery*. Портфоліо *Discovery* надає набір готових до застосування аналітичних функцій *SQL*-аналізу, *SQL-MapReduce®* і *Graph*, функції часових рядів, статистичних методів, зокрема прогнозного аналізу, аналітики тексту та багато іншого для дослідження *BigData*. Містить модулі [34] збирання та підготовки даних, засобів аналітики та візуалізації. Перший із них забезпечує доступ до мультиструктурованих даних, що зберігаються в *Apache™ Hadoop™*, *Teradata Data Warehouse* та інших реляційних СУБД (*RDBMS*). Модуль підготовки даних містить адаптери та перетворювачі, які дають змогу аналізувати та інтерпретувати вміст блогів, *XML*-документів, електронних листів і журналів пристроїв. Пропонований набір функцій надає такі можливості: фільтри для видалення викидів значень даних; *Apache Log Parser* для підтримки форматів журналів для користувача з певним користувачем веб-серверів *Apache; XML* і *JSON Parsers* для аналізування та підготовки *XML*-журналів, що створюються такими застосунками як веб-журнали та касові журнали в роздрібних магазинах тощо. Функції перетворення даних передбачають розпакування форматів для перетворення складних, неструктурованих даних у значущі формати для аналітики.

Платформу *Teradata Aster Discovery* спроєктовано для даних, поєднаних терміном *Big-Data*. Це гарантує застосування всіх доступних даних від різних джерел (веб-журнали, сенсорні мережі, соціальні медіа, докладні дані, отримані від телекомунікаційних мереж та засобів Інтернету речей (*Internet of Things*), астрономічні та військові спостереження, біологічні системи, медичні записи, фото- та відеоархіви), з можливістю розміщення їх безмежних обсягів та надання необхідних відомостей. Синергетичні мультижанрові засоби аналітики застосовують множинні механізми (*SQL, MapReduce*, статистичні, аналіз тексту та *Graph*-аналіз) для виявлення нових ідей. Розширена бібліотека попередньо побудованих *SQL, SQL-MapReduce*, функцій *Graph* дає змогу розпочати збирання, підготовку, аналіз та візуалізацію даних однією дією. Крім цього, БД *Teradata* та *Teradata Aster Discovery Platform* є доступними як служби для сховищ даних та аналітики виявлення інформації в даних. Єдина архітектура даних *UDA Teradata™* для всіх типів даних забезпечує можливість хмарної доставки.

Data Preparation (підготовка даних) – це процес їх об'єднання, приведення до єдиного формату та очищення з метою подальшого аналізування та вирішення інших бізнес-завдань [35]. У формулюванні *Gartner* – це ітеративно-гнучкий процес пошуку, об'єднання, очищення та трансформації первинних, необроблених даних (*raw data*) в набори керованих (*data curation*) даних для інтеграції самообслуговуванням, аналізу даних, виявлення інформації в даних для бізнес-аналітики, або, інакше, в цільові відібрані набори, що їх застосовують в інтелектуальних дослідженнях даних [36].

Data Curation – це засіб керування даними, що робить їх кориснішими для користувачів, які займаються виявленням та аналізуванням даних. Для цього дані збираються з різних джерел, потім їх інтегрують у репозиторії, де вони набувають вищої цінності, ніж незалежні частини.

Оцінювання якості даних (*Data quality verification*) – необхідний етап попередньої підготовки даних, суть якого полягає у виявленні проблем, які не дають змоги коректно аналізувати дані й знижують значущість і достовірність результатів аналізування.

Очищення даних (*Data clearing*) – це підвищення якості даних за допомогою усунення

виявлених проблем, таких як порушення структури, цілісності та повноти, некоректні формати, фіктивні значення, пропуски, суперечності, дублікати.

Попереднє оброблення даних (*Data preprocessing*) – це процес підвищення якості даних за допомогою застосування аналітичних методів для очищення від шумів і згладжування рядів даних, зниження розмірності вхідних даних, усунення незначущих ознак.

Трансформація даних (*Data transformation*) – це перетворення даних до певного подання, формату або виду, оптимального з погляду конкретного методу аналізування.

Інтеграція даних (*Data Integration*) дає змогу застосовувати дані з різних джерел незалежно від їхнього формату. Розрізняють технології *ETL* (вилучення, трансформація, завантаження), а також технологія віртуалізації даних, тобто доступ до даних без безпосереднього передавання їх по каналах зв'язку [34].

Протягом багатьох років наявними є інструментарії підготовки даних, які допомагають готувати реляційні дані до видобування інформації з них (*data mining*). Їх переважно було вбудовано в продукти інтелектуального оброблення даних або тісно інтегровано з такими застосунками. Зараз з'являється клас автономних розробок, що пропонують однаковий тип можливостей для всіх типів (структурованих, напівструктурованих та неструктурованих) та джерел даних; ці розробки зорієнтовано насамперед на бізнес-аналітиків. Ключовим компонентом таких розробок є їхня здатність надавати можливості самообслуговування, що дають змогу досвідченим користувачам з навичками бізнес-аналітики без залучення фахівців *IT*-служб об'єднувати, перетворювати та «очищати» відповідні дані перед аналізом, тобто «готувати» їх до інтелектуального оброблення. Отримання адекватної картини бізнесу потребує забезпечення високої якості даних, що їх аналізують, тобто, необхідно застосовувати розширений набір засобів для очищення даних: виявлення неповних або помилкових даних, інформації, що дублюється, приведення даних із різних джерел до єдиного формату. У цьому класі більшість інструментів націлено на бізнес-аналітиків, але є продукти, орієнтовані значною мірою на фахівців з даних. Скільки і яких знань потрібно користувачам, залежить від конкретної ситуації [37].

Self Service Data: епоха самообслуговування настала (*Gartner*, 2016)[12]

Високооплачувані дослідники даних витрачають 80% свого часу на виявлення та «підготування» даних, перш ніж запускати на них свої моделі. Це є неймовірно неефективним, і невід'ємна затримка від гіпотези до розуміння щороку коштує компаніям великих витрат. Традиційні *IT*-інструменти, призначені для підтримання запитів даних бізнес-підрозділів, не встигають за збільшенням нових джерел даних, багато з яких є напівструктурованими. Бізнес-аналітика самообслуговування була серед побажань компа-ніям протягом тривалого часу, і вона ще досі є пріоритетною. Дедалі більший попит є результатом потреби бізнес-користувачів у більшій гнучкості та самостійності у звітності й аналізі. Самообслуговування *BI* – це тренд із трохи розпливчастим визначенням. У найзагальнішому розумінні завдання самообслуговування *BI* – це ті, які бізнес-користувачі виконують самі, а не передають їх *IT* для виконання. Для кожної ролі *BI* самообслуговування може допомогти користувачам виконувати різні завдання. Потреба в самообслуговуванні в середовищі *BI* залежить від вимог користувача.

Термін «дані самообслуговування» означає можливість для бізнес-користувача, з метою поліпшення розуміння бізнесу, самостійно знаходити та застосовувати будь-які корпоративні дані, щоб отримати потрібну інформацію про бізнес. Користувачеві, щоб подолати проблеми з даними та збільшити

[12] *Swoyer Steve*, February 17, 2016. Gartner Makes it Official: The Age of Self-Service Data is Upon Us – https://tdwi.org/articles/2016/02/17/age-of-self-service.aspx.

прибуток або зменшити витрати, необхідно надати набір інтегрованих інструментів самообслуговування даних. Користувач повинен мати можливість самостійно виявити набір даних у каталозі, потім призначити його інструментарію підготовки для очищення або інтегрування локальних даних із різних джерел з іншими наборами даних. Тобто, аналітика самообслуговування – це тип бізнес-аналітики, який дає змогу користувачам отримувати доступ до даних й аналізувати їх, не покладаючись на підтримку спеціалістів *IT* або *BI* [38, 39].

Наведімо розробку *Greenplum* [40], що демонструє принцип розв'язання проблеми самообслуговування даних. *Greenplum Chorus*™ – це клас програмних засобів та користувацький інтерфейс для побудови запитів і візуалізації даних із додаванням рівня соціальних мереж до аналізу набору даних. Спрощуючи взаємодію спеціалістів різного профілю та загальний складний аналітичний процес, *Chorus* (проєкт із відкритим кодом) дає користувачам змогу взаємодіяти з наборами даних і створювати власні «пісочниці» для тестування та обговорення отриманих результатів у спільній роботі в стилі *Facebook*. Можна завантажити початковий код і почати роботу, змінити та поширити його на будь-яке середовище. Це також дає змогу створити екосистеми застосунків та стартапів у середовищі *Big Data*, що веде до розширення можливостей продуктів із більшою ефективністю, ніж цього могли би досягти дослідники самостійно. *Chorus* містить проєкти від таких партнерів, як *Squid Solutions, Gnip, Kaggle* і *Tableau* та діє як спільна платформа досліджень даних, за допомогою якої користувачі отримують прямий доступ до каналу *Twitter* від *Gnip*, виконують розширений аналіз за менший час, залучають експертів *Kaggle* за замовленням та обмінюються безперешкодно інформацією за допомогою розширених візуалізацій *Tableau* [41]. Платформа є доступною не лише для фахівців, а й для керівників бізнес-підрозділів, а також для пересічних співробітників. Інноваційні функції послуг самообслуговування, що дають фахівцям змогу легше взаємодіяти та отримувати необхідні знання про дані підприємства. *Greenplum Software*, провідна світова компанія у сфері інфраструктури даних, представила на початку 2010 р. розробку *Greenplum Chorus*, першої комерційно наявної платформи *Enterprise Data Cloud* [42]. *Chorus* тоді обслуговував потреби аналітиків даних, *IT* та бізнес-керівників в обсязі трьох основних функцій:

– надання послуг самообслуговування вітрин даних (*data marts*) і захищених контрольованих «пісочниць» (*sandboxes*), зменшуючи *IT*-витрати на керування та операційну складність, зазвичай пов'язані з розгортанням масиву даних;

– створення служби даних, що дає аналітикам змогу виявляти, об'єднувати та ділитися корисними наборами даних і всередині вітрин, й із сусідніми вітринами;

– спільна робота з даними, що надається великими можливостями соціальних мереж, які пов'язують дані й аналітиків, щоб пришвидшити співпрацю та розуміння.

Платформи підготовки даних на засадах самообслуговування призначені забезпечувати своїх користувачів функційними можливостями підготовки даних до аналізування: очищення даних, видалення або заповнення порожніх полів, видалення дублікатів даних, додавання даних з інших джерел (зокрема з ідентифікацією ключа зв'язку), перетворення формату даних, а також згортання й агрегування даних.

Самообслуговування даних було виконано на основі програмних стартапів та розробок кількох великих компаній. Отже, є чимало інструментарію, який вирішує конкретну проблему конвеєра самообслуговування даних (рис. 3). Завданням стало керування цим громіздким Франкенштейном середовища даних, яке було створено з метою спрощення доступу до даних та розумінням потреб бізнес-користувача [39].

Засоби підготовки даних на засадах самообслуговування (див. рис. 3) відповідають

2018 року [43, 44]. Найбільш функційно розвинені продукти розташовано ближче до центру діаграми. Серед них виокремлюють три категорії: стартапи, які створили окрему платформу для підготовки даних (*Paxata, Trifacta, Tamr* і *Alation*); постачальники засобів інтеграції даних та/або рішень, що поліпшують якість даних (процеси стандартизації, очищення тощо), які розширили наявні можливості в цій галузі (*Informatica, Progress* і *IBM*); і постачальники засобів бізнес-аналітики/аналітики (знову *IBM, Rocket Software, Alteryx, ClearStory* і *Datawatch*) – те, що деякі продукти вбудовано в ширші аналітичні пропозиції, не пов'язує підготовку даних з аналітикою як такою. Помітно відсутні постачальники засобів баз даних/сховищ, такі як *Teradata* [34], *HP* (*Vertica*) [45], *Pivotal* [40] тощо. У зв'язку зі збільшенням кількості постачальників зроблено висновок, що компанії у сфері бізнес-аналітики, аналітики та сховищ даних, які не мають засобів власної підготовки даних, почнуть програвати конкурентам із такою можливістю.

Етап змішування необхідних даних із різних джерел пов'язано з пошуком або розвідкою та ідентифікацією доступних даних, що стосуються розглянутого процесу (об'єкта). Це розширення для підготовки даних є еквівалентним виявленню даних. Однак цей термін уже застосовують, тому замість нього введено поняття каталогізації даних [43]. Каталогізація даних означає автоматичний обхід доступних джерел даних, збір метаданих про ці джерела та їхній вміст, щоб дати змогу користувачам виконувати пошук за каталогом і знайти необхідні дані.

Каталог даних [44] – це сховище інформації про активи даних компанії: які дані зберігаються, у якому форматі їх зберігають, у яких бізнес-доменах (областях) ці дані є актуальними та де вони містяться (в яких базах даних та/або файлах). Інформація в каталозі даних може бути додатково класифікована за географічною ознакою, часом, керування доступом тощо. Каталоги даних індексуються та є доступними для пошуку, що зменшує фактор часу; можуть охоплювати різні джерела, зокрема соціальні мережі, інтернет-ресурси, складні текстові документи тощо, допомагають усунути їхню розрізненість; забезпечують самообслуговування та, отже, підвищують продуктивність внаслідок того, що дають змогу бізнес-користувачам шукати цікаву для них інформацію, не вдаючись до послуг *IT*. Нарешті, з огляду на потік даних, якими перевантажено багато компаній, каталоги допомагають у цьому середовищі, щоб користувачі могли бачити, які дані є актуальними, а які ні. Створення каталогів даних може бути аналогічне тому самому *Google*, що надає «каталог» веб-документів, застосовуючи веб-павуки або інші технології для створення зручних можливостей пошуку.

Варто зазначити, що назріває серйозна проблема розрізнених сховищ. Великі організації неминуче матимуть кілька каталогів даних замість єдиного каталогу, що охоплює все підприємство і який застосовує усі відповідні продукти та інструменти. Доцільніше [44] застосовувати граф знань – це графічні уявлення зв'язків між речами. У світі великих даних та *IT* застосування графів знань як мета-рівня, що охоплює каталоги даних, має великий потенціал. *Cambridge Semantics* фактично робить це в рамках свого продукту *Anzo*, підтримуючи концепцію підкаталогів. У цьому контексті варто зазначити, що низка постачальників каталогів пропонують можливості графа знань: наприклад, це роблять і *IBM,* і *Informatica*.

Ключові можливості сучасних засобів підготовки даних

Ключові властивості [36] оцінювалися за категоріями функціональності, наведеними далі.

1. Доступ до даних і профілювання даних (*Data Profiling*) [35,46], фільтрація та перетворення. Це стосується можливості отримати доступ та інтегрувати дані різних типів із розрізнених джерел, а також для перетворення та підготовки їх для моделювання. Профілювання даних – це ключ до їхнього розуміння, оскільки воно

забезпечує високорівневу статистику якості даних (характер розподілу величин, наявність викидів, параметри вибірки; кількість рядків, типи даних у стовпчиках, мінімальні, максимальні та середні значення по стовпцях, кількість нульових значень тощо або резюме про ці дані). Також профілювання полегшує вибір одного з кількох відповідних наборів даних, тобто це процес вивчення даних, доступних з наявних джерел інформації. Ці знання потрібні для оцінювання якості даних (пошук пропущених значень, порушення цілісності та бізнес-логіки зв'язків між значеннями полів таблиць, відповідність даних певним стандартам або критеріям тощо) в рамках їх підготовки до інтелектуального аналізування; оцінювання ризику, пов'язаного з інтеграцією даних у нові застосунки; які метадані вихідної інформації; які поля будуть ключовими, охоплюючи визначення зовнішніх ключів і функціональних залежностей; організації процесів роботи з бункерами (*Data Silos*) і озерами даних (*Data Lakes* [45]), корпоративними сховищами даних (КСД) і розробки *ETL /ELT*-процесів тощо.

2. Каталогізація даних і базове керування метаданими. Для ефективного пошуку даних, найпридатніших для вирішення конкретного завдання, підтримується створення та пошук метаданих, каталогізація джерел даних, перетворень, дій користувача стосовно джерела даних, атрибутів джерела даних, походження даних і відношень, а також *API*-інтерфейси для забезпечення доступу до каталогу метаданих (тобто репозиторію метаданих). У каталозі зберігається також описовий індекс, який вказує на місце розташування доступних даних. Внаслідок застосування аналітики необроблених даних, моделі створюються та генеруються знизу вгору, а не (їх розробляють) згори вниз. Це безперервний процес накопичення метаданих, що ґрунтується на фактичному застосуванні даних. *Gartner* [36] зазначає, що деякі інструменти підготовки даних охоплюють каталогізацію даних у робочий процес їхньої підготовки, і ці спеціалізовані каталогізації даних є точковими рішеннями, що дають змогу користувачам здійснювати інвентаризацію активів даних, інтегрованих цими рішеннями.

3. Структурування, моделювання та перетворення даних підтримують змішування та гармонізацію[13] даних; очищення даних; фільтрацію і призначені користувачу для обчислення, групування та ієрархії. Це охоплює гнучке моделювання[14] та структурування даних, яке дає користувачам змогу зазначати типи даних і зв'язки між ними. Прогресивніші можливості автоматично встановлюють походження даних або роблять висновок про їхню структуру з джерела даних і генерують семантичні моделі та онтології, такі як логічні моделі даних і *Hive*-схеми[15] . Очищення застосовують для оцінювання якості даних і саме тому воно є невід'ємною частиною їхньої підготовки. Воно зазвичай охоплює перевірку, вилучення повторень і збагачення даних.

4. Базова якість та безпека даних. Інтеграція з інструментами, що підтримують керування та впорядкування інформації, а також можливості шифрування даних, дозволів користувачів та передавання даних. Це також охоплює функції безпеки, такі як маскування даних, автентифікація платформи та фільтрація безпеки на рівні користувача/групи/ролі, а також через інтеграцію з корпоративними системами *LDAP* (*Lightweight Directory Access Protocol*, полегшений протокол доступу до каталогів) та/або системи *Active Directory, SSO* (технологія єдиного входу), успадкування безпеки вихідної системи, безпека на рівні рядків і стовпців, а також ведення журналу,

[13] Гармонізація – приведення даних до взаємної відповідності, наприклад, міжнародного стандарту обміну даними *SDMX*.

[14] Якщо коротко, то моделювання даних – це процес створення моделі даних для зберігання їх у базі даних, яка є концептуальним уявленням про об'єкти даних, зв'язки між ними та правила

[15] *Hive* є системою під час читання – *Schema on Read,* тоді як *RDBMS* (*Relational Database Management System*) зазвичай є системами під час записування – *Schema on Write*. Традиційні *RDBMS* перевіряють схему під час записування даних. Якщо дані не відповідають структурі, їх відхиляють. *Hive* не дбає про структуру бази даних і не перевіряє схему під час завантаження даних у базу. Навпаки, схема перевіряється лише після виконання запиту до бази даних [47].

моніторинг використання даних та активів. Деякі інструменти пропонують візуалізацію в процесі підготовки даних, щоб показати розподіл даних, сторонні дані та інші взаємозв'язки.

5. Збагачення даних. Підтримка основних можливостей поповнення даних, зокрема виявлення сутностей (захоплення атрибутів даних, інтегрованих за допомогою інструмента підготовки даних); розроблення атрибутів (що дає змогу спеціалістам із процесів аналізу вдосконалювати набори атрибутів для інтегрованих даних на основі вимог галузі); та покращення циклу збагачення даних за допомогою машинного навчання для майбутнього застосування з додаванням нових наборів даних для підвищення продуктивності аналітиків та інших користувачів.

6. Співпраця з користувачами. Спрощує обмін запитами та наборами даних, спільне застосування та розвиток моделей. При цьому важливо, щоб інструменти самообслуговування не було ізольовано. Їм необхідно мати можливість каталогізувати, спільно застосовувати та керувати метаданими або за допомогою вбудованих функцій каталогізації даних, або за допомогою їх здатності обмінюватися метаданими за допомогою повніших та ширших інструментів керування метаданими. Вони охоплюють інші важливі можливості, крім каталогізації даних, такі як походження даних, бізнес-глосарій, керування правилами, аналіз впливу та семантичні структури, які дають *IT*-спеціалістам змогу керувати та перевіряти потік підготовки даних стосовно якості та керування перед їх розвитком.

7. Доступ до джерел даних та можливості підключення. *API*-інтерфейси та підключення ґрунтуються на стандартах (*standarts-based*), зокрема нативний (власний) доступ до хмарних програм і джерел даних: локальних, корпоративних, реляційних та неструктурованих, *NoSQL*, *Hadoop* і різних форматів файлів (*XML, JSON, .csv*), а також нативний доступ до ретельно (вищої якості) підібраних даних. Виробники сучасних засобів підготовки даних також додають у свій портфель підтримку потокового передавання даних, таких як дані про машини та дані Інтернету речей (*IoT*).

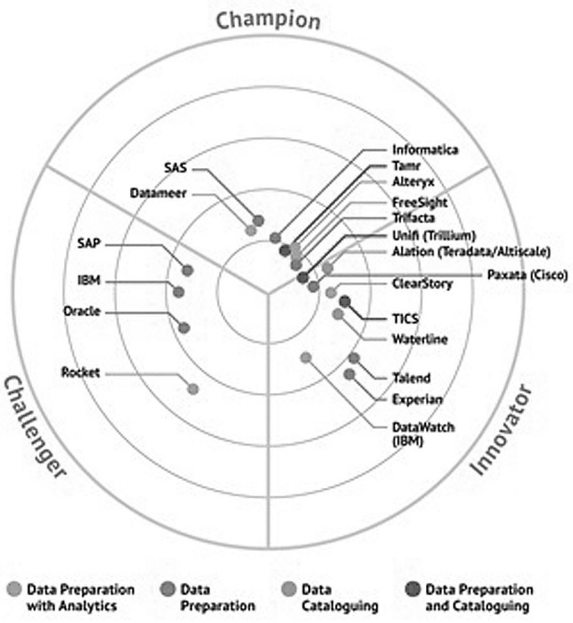

Рис. 3. Платформи підготовки даних на засадах самообслуговування

8. Машинне навчання. Застосування машинного навчання та штучного інтелекту для покращення й автоматизації процесу підготовки даних. Деякі інструменти надають алгоритми, що дають користувачам змогу ідентифікувати структури даних, схеми та зв'язки з різними рівнями деталізації, а також можливість структурувати набори даних за початкового їх введення.

Машинне навчання також дає користувачам змогу зрозуміти, коли вони можуть задіяти свої потоки та направити їх в інтеграційні потоки даних по всій організації внаслідок можливості обмінюватися метаданими у зростаючий двонаправлений спосіб із зовнішніми інструментами керування метаданими, якістю даних та керування.

9. Моделі розгортання. Ці інструменти можуть бути розгорнуті локально чи у хмарі або спільно. Останній, гібридний підхід дає змогу користувачам залишати дані на місці оброблення, а не переміщувати їх на платформу

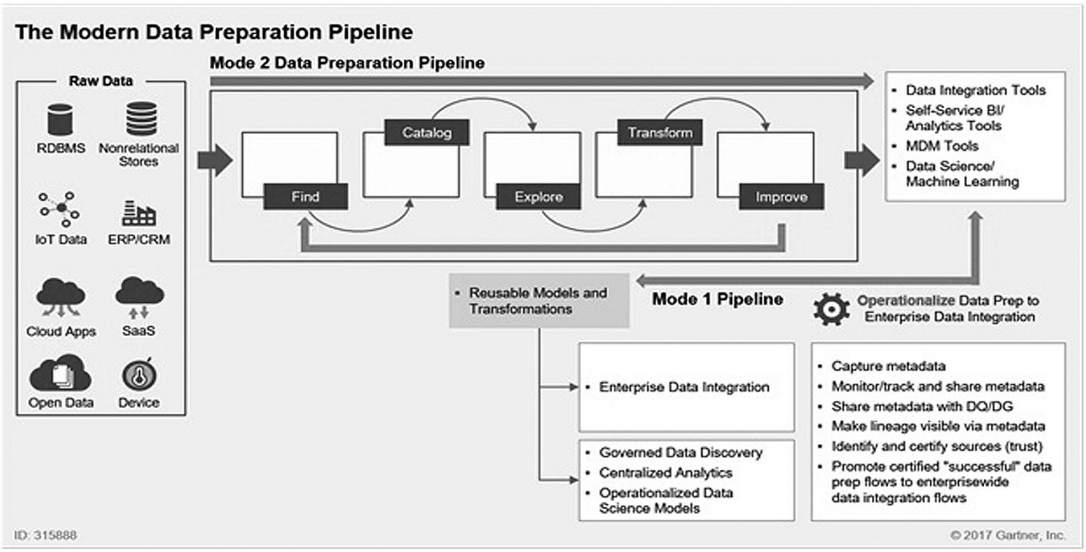

Рис. 4. Операціоналізація підготовки даних – сучасний конвеєр підготовки даних

самообслуговування для підготовки даних [34]. Незначна кількість постачальників також підтримує розгортання з кількома хмарами.

10. Пропозиції або шаблони для конкретної області (домена) або зрізу. Упаковані (пакетовані) шаблони або пропозиції для даних і моделей, що сприяють пришвидшенню підготовки даних та розумінню їх. Це стає особливо корисним для складних у застосуванні синдикованих (об'єднаних у синдикат) наборів даних.

11. Об'єднання з іншими інтегрованими платформами та рішеннями щодо керування інформацією. Можливість інтеграції узгоджених наборів даних із платформами інтеграції та аналітики/бізнес-аналітики та оброблення даних через *API*, веб-сервіси або вбудовану підтримку форматів файлів для партнерів (наприклад, *.tde* для *Tableau*, *.qvd* для *Qlik* та *.pbi* для *Microsoft Power BI*). Водночас інформація, отримана в результаті роботи з виявлення інформації в даних, може бути використана в процесі керування бізнес-дисципліною та розподілом у разі потреби виявлених даних та метаданих відповідно до вироблених рішень.

Для підтримки постійно змінюваних вимог організації щодо варіантів застосування, сучасні засоби мають підтримувати та забезпечувати високорівневі можливості підготовки основних даних (*Master Data Management* [16]*, MDM*) [48].

Підготовка даних у сучасному виробництві

Операціоналізацію[17] потоків підготовки даних показано на рис. 4. У ньому відображено практику керування двома окремими узгодженими способами надання *IT*-послуг. Це режим [49], коли один з них зосереджено на передбачуваності, інший – на дослідженні. Режим 1 припускає стабільність і є оптимізованим для очікуваніших і зрозуміліших областей. Режим 2 – гнучкий, експериментальний для вирішення нових проблем і оптимізований для областей невизначеності. Ініціатива в цьому разі часто починається з

[16] *Майстер-дані* – це дані про ідентифіковувані сутності, в яких укладено ключову інформацію про бізнес загалом (технології, матеріали, продукти, клієнти тощо), спільно застосовувані у взаємодійових інформаційних системах.

[17] *Операціоналізація* – це процес визначення виміру явища, яке не можна виміряти безпосередньо, хоча його існування визначається іншими явищами. Таким чином, операціоналізація визначає нечітку концепцію, щоб зробити її чітко помітною, вимірною та зрозумілою за допомогою емпіричного спостереження, інакше перетворення теоретичного судження з метою його емпіричної перевірки – *https://en.wikipedia.org/wiki/Operationalization*.

гіпотези, яка перевіряється та адаптується в процесі, що передбачає короткі ітерації, тут потенційно застосовують підхід мінімально життєздатного продукту (*minimum viable product, MVP*). Поєднання передбачуванішої еволюції продуктів і технологій (Режим 1) з новими та інноваційними (Режим 2) є суттю бімодальних можливостей підприємства. Обидва режими відіграють важливу роль у цифровий трансформації [49, 50].

Більшість завдань з підготовки даних починаються з виконання Режиму 2 в конвеєрі підготовки даних (див. рис. 4), коли бізнес-користувачі підключаються до кількох, часто мінливих, джерел даних і коли часу для осмислення ситуації замало. Ці завдання виникають через бізнес-необхідність експериментувати з новими джерелами й типами даних, коли бізнес не впевнений у кінцевому успіху, життєздатності експерименту й хоче швидко перевірити гіпотезу, не чекаючи підтримки *IT*-відділу. Саме тут можливості самообслуговування інструментів підготовки даних дають бізнесу змогу значно скоротити час прийняття рішення.

У разі успішного завершення експерименту в Режимі 2, бізнес реалізує отримані результати. Для цього він забезпечує впровадження *IT*-фахівцями необхідних «пісочниць» і процесів розвитку на основі стратегій і правил керування. *IT*-відділ надає можливості для збирання та обміну метаданими з застосовуваних потоків підготовки даних у двох напрямках (див. рис. 3). Також встановлюються інструменти керування даними, такі як керування метаданими, якість даних, інтеграція і керування даними, щоб відстежувати й аналізувати повторне їх застосування, а потім рекомендувати, а також візуалізувати розвиток цих потоків підготовки даних у Режимі 2 у ширший режим багаторазового застосування перетворень для Режиму 1 через інтеграцію даних підприємства. Цей процес гарантує, що організація зможе розвивати ініціативи, які починалися як «самообслуговування», для керування виявленням інформації в даних, централізованою аналітикою, операціоналізацією даних та іншими критично важливими ініціативами, коли це необхідно. Важливо, щоб інструменти підготовки даних підтримували «операціоналізацію потоківідготовки даних», а також випадки їх застосування на підприємстві.

Тенденції, сприятливі до впровадження інструментарію підготовки даних – це еволюція від даних самообслуговування до індустріальної підготовки даних; підготовка даних та машинне навчання; конвергенція. Перша з них зумовлена тим, що ринок інструментарію для підготовки даних розвивається від самообслуговування до індустрії аналогічно до ринку для сучасних платформ *BI* та аналітики. Спочатку інструментарій, через його гнучкість та простоту застосування, почали застосовувати для аналітики самообслуговування та дослідників даних, щоб пришвидшити їх підготовку для інтерактивного аналізу та *data science*. Можливості підготовки даних на засадах самообслуговування тепер доступні як ключова вбудована функція сучасних платформ *BIA*, платформ *data science* і машинного навчання й інструментарію для інтеграції даних. Однак згодом спеціальні автономні інструментарії постали перед необхідністю створювати надійні складні моделі даних зі все більшої кількості наборів даних, що застосовують різні інструменти. Це виходить за рамки можливостей більшості вбудованих систем підготовки даних.

Автономні інструментарії тепер дають змогу інженерам та аналітикам даних у централізованих командах швидко готувати контрольовані та масштабовані набори даних для застосування в рамках усього підприємства. Це сприяє передаванню деякої відповідальності за бізнес-орієнтовані завдання інтеграції менш кваліфікованим *IT*-спеціалістам або навіть недостатньо обізнаним з цією областю, що дає змогу фахівцям з даних та аналітики вивільнити час на вирішення складних завдань.

Підготовка даних та машинне навчання. Мірою того, як дані стають складнішими, практично неможливо або надто довго ке-

рувати, очищати, гармонізувати та формувати дані вручну. Машинне навчання стало критично важливою функцією для реалізації та автоматизації трудомістких і схильних до помилок завдань щодо підготовки даних. Більшість постачальників інструментарію для підготовки даних уже долучили деякі алгоритми машинного навчання до своїх інструментів підготовки даних, щоб покращити та зробити цей процес продуктивнішим.

Конвергенція. Спостерігається дедалі більше зближення можливостей для підготовки даних з інструментами виявлення зв'язків і дозволу сутностей, які можуть підтримувати об'єднання та консолідацію даних у висхідному напрямку (наприклад, *Tamr* та *Reltio*), інструментами, що підтримують профіль даних та можливості оцінювання якості даних (наприклад, *Informatica* та *Talend*), можливості інтеграції потокових даних (наприклад, *Striim* та *StreamSets*), каталогізації даних (наприклад, *Alation* та *Waterline Data*). Усі ці дії є важливою частиною загального конвеєра підготовки даних і додають функції іншого рівня, або набувають їх для створення повнішого робочого процесу (наприклад, *Alteryx, Unifi, Trifacta, Paxata, Datameer* та *Datawatch*).

Висновки

Загалом підготовка даних – найбільш часовитратна задача в аналітиці та бізнес-аналітиці – перетворюється від самообслуговування до корпоративного завдання.

Ринок для підготовки даних перетворився з інструментів, що підтримують лише сценарії самообслуговування, на платформи, що дають змогу співробітникам, які працюють з даними та аналітикою, створювати гнучкі набори даних із можливістю пошуку контенту, розподіленого в масштабі підприємств. Більшість пропозицій постачальників підтримують профілювання даних, їхні дослідження, перетворення, моделювання та контролювання, а також підтримку метаданих. Більшість (80 відсотків) постачальників вбудовують деякі функції каталогізації даних і пропонують різні рівні можливостей машинного навчання.

Віце-президент із досліджень та відомий аналітик *Gartner* – Тед Фрідман (*Ted Friedman*), голова саміту *Gartner Data & Analytics Summit* звернув увагу на зближення ролей та інтересів у галузі даних та аналітики[18]. Організації дедалі частіше створюють єдину групу з обробки даних та аналітики на чолі з головним співробітником із обробки даних (*chief data officer, CDO*). 25 відсотків великих світових організацій уже найняли *CDO*.

Висловлено припущення, що доповнюване машинним навчанням керування основними даними (*MDM*), якість даних, підготовка даних і каталоги даних об'єднаються в єдину сучасну платформу керування корпоративною інформацією (*Enterprise Information Management, EIM*), застосовну для більшості нових аналітичних проєктів.

Якісне вирішення зазначених у цьому огляді проблем першого рівня роботи з даними забезпечить у сучасних аналітичних платформах дослідження даних успішне проникнення в їхню суть на рівні отримання знань за допомогою машинного навчання та штучного інтелекту. Це дасть змогу прогнозувати майбутні результати у керованих об'єктів (процесах) та прийняття обґрунтованих рішень.

Матеріал огляду побудовано здебільшого на рішеннях, що стосуються бізнес-інтелекту, призначених для задач із корпоративними даними. Але всі розглянуті в роботі основні аспекти роботи з даними, застосовуються й на платформах обробки даних (*Data Science Platform*). Багато постачальників *BI* розширили можливості своїх систем для виконання досконалішої аналітики, включно з *Data Science*. Словосполучення «*Data Science*» вони додали до своїх маркетингових досліджень, а термін «просунута аналітика» трохи втратив популярність. Стосовно корпоративних даних. Аналітична платформа *Data Science Platform* використовує розподілену архітектуру, масовий паралелізм обробки, віртуалізацію

[18] *Egham, U.K., December 16, 2016. The Rise of the Chief Data Officer Signals an Age of Infinite Possibilities – https://www.gartner.com/en/newsroom/press-releases/2016-12-16-rise-of-the-chief-data-officer-signals-age-of-infinite-possibilities.*

даних, обчислення у пам'яті тощо. Поєднання традиційної обробки реляційних даних з обчисленнями на широко відомій програмній інфраструктурі *Apache Hadoop*, у яку інтегровано низку ком-понентів екосистеми *Hadoop* (*Apache Hive, HBase, Spark, Solr* тощо) з необхідними цільо-вими функціями, уможливлює створення повнофункціональної платформи зберігання та обробки структурованих і неструктурованих даних.

REFERENCES

1. Gritsenko, V.I., Oursatyev, A.A., 2017. "*Big Data* and Tools for Analytics". Upravlausie sistemy i masiny, 4, pp. 3-14. (In Russian).
2. Ginsburg, J., Mohebbi, M., Patel, R. et al. Detecting influenza epidemics using search engine query data. *Nature*. 2009, 457, pp. 1012–1014, http://www.nature.com/nature/journal/ v457/n7232/full/nature07634.html.
3. Gritsenko, V.I., Oursatyev, A.A., 2011. "Information Technologies: the Tendency, the Ways of the Development". Upravlausie sistemy i masiny, 5, pp. 3-20. (In Russian).
4. Luhn, H.P., 1958. "A *Business Intelligence* System". IBM Journal of Research and Development, Vol. 2, Issue 4, pp. 314–319.
5. Martens, Ch. The maturing of BI. Interview: Hyperion chief strategy officer Howard Dresner discusses how as BI matures, companies should too. InfoWorld. Sep. 22. 2006. [online]. Available at: <https://www.infoworld.com/article/2661157/database/the-maturing-of-bi.html> [Accessed: 17 Dec. 2021].
6. Laurent Duval. Original Meaning of "Intelligence" in "Business Intelligence". [online]. Available at: <https://datascience.stackexchange.com/questions/8016/original-meaning-of-intelligence-in-business-intelligence> 07 Nov. 2015].
7. What's the Difference Between Business Intelligence (BI) and EPM? [online]. Available at: < http://blog.hostanalytics.com/whats-the-difference-between-business-intelligence-bi-and-epm> [Accessed: 07 Nov. 2021].
8. Dresner, H. Predicts the future of business intelligence. [online]. Available at: <http://searchbusinessanalytics.techtarget.com/podcast/Howard-Dresner-predicts-the-future-of-business-intelligence> [Accessed: 17 Dec. 2021].
9. Schlegel, K., Sallam, R.L., Yuen, D. et al. Magic Quadrant for Business Intelligence and Analytics Platforms. [online]. Available at: < http://business-view.dk/wp-content/uploads/2015/02/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms-ALL.pdf> [Accessed: 05 Feb. 2013].
10. Herschel, G., Linden, A., Kart, L. Magic Quadrant for Advanced Analytics Platforms. [online]. Available at: <https://pdfs.semanticscholar.org /1a9f /ff52e8084d0da00491e54d45113bd81d2e91.pdf> [Accessed: 19 Feb. 2014].
11. Herschel, G., Linden, A., Kart, L. Magic Quadrant for Advanced Analytics Platforms. [online]. Available at: <https://davidhoglund.typepad.com/files/magic-quadrant-for-advanced-analytics-platforms.pdf> [Accessed: 17 Feb. 2015].
12. Magic Quadrant for Business Intelligence and Analytics Platforms / R.L. Sallam, J. Tapadinhas, J. Parenteau et al. [online]. Available at: <http://www.thgcfo.com/wp-content/uploads/2014/02/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms.pdf> [Accessed: 20 Feb. 2014].
13. Magic Quadrant for Business Intelligence and Analytics Platforms / Rita L. Sallam, Bill Hostmann, Kurt Schlegel et al. [online]. Available at: <http://zzircon.com/wp-content/uploads/2015/04/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms-2015.pdf> [Accessed: 23 Feb. 2015].
14. Business intelligence systems 2013. Market overview. TAdviser. [online]. Available at: <http://www.old.rbcgrp.com/files/QlikView_TAdviser2013.pdf> [Accessed: 17 Dec. 2021].
15. Dinsmore, T.W. Disruptive Analytics: Charting Your Strategy for Next-Generation Business Analytics. Apress, 2016, 262 p, https://www.apress.com/us/book/9781484213124> [Accessed: 17 Dec. 2021].
16. Nikolaev, O. Gartner: analytics should become a top priority for business. [online]. Available at: <http://channel4it.com/publications/Gartner-analitika-dolzhna-stat-glavnym-prioritetom-dlya-biznesa-5204.html#> 17 Oct. 2014].
17. Chabrier, A. Data Types for Data Sciences. [online]. Available at: <https://towardsdatascience.com/data-types-for-data-sciences-65dcbda6177c1818> [Accessed: 17 Dec. 2021].
18. Prohnozna analityka vid SAP – SAP. Predictive Analytics. [online]. Available at: <https://jetbi.ru/obzor-sap-predictive-analytics> [Accessed: 23 Aug. 2018].
19. Predictive analytics. [online] Available at:<https://en.wikipedia.org/wiki/Predictive_ analytics> 23 Aug. 2018].
20. Stepashko, V.S., Yefimenko, S.N, 2018. "Predictive Analytics as an effective tool for decision support in Digital Economics Systems". Upravlausie sistemy i masiny, 6, pp. 25-35. (In Ukraine).
21. Stepashko, V.S., 2017. "The Achievements and Prospects of Inductive Modeling". Upravlausie sistemy i masiny, 2, pp. 58-73. (In Russian).
22. Descriptive, Predictive, and Prescriptive Analytics Explained. [online] Available at:<https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/> [Accessed 05 Jun. 2019].

23. Prescriptive Analytics. [online] Available at:<https://en.wikipedia.org/ wiki/Prescriptive_analytics> 07 May 2019].
24. Predictive Analytics vs. Prescriptive Analytics: What Is the Difference? Available at:<https://www.proponent.com/predictive-analytics-vs-prescriptive-analytics/> [Accessed 07 May 2019].
25. Descriptive, Predictive and Prescriptive Analytics. Available at:<http://www.gurobi.com/resources/prescriptive-analytics>[Accessed 07 May 2019].
26. IBM Analytics. Prescriptive analytics. [online] Available at:<https://www.ibm.com/ analytics/prescriptive-analytics> [Accessed 07 May 2019].
27. Frankenfield, J., Prescriptive Analytics. [online] Available at: <https://www.investopedia.com/terms/p/prescriptive-analytics.asp> [Accessed 06 March 2019].
28. Schlegel, K., 2008. The Rise of Data Discovery Tools. [online] Available at:<https://www.gartner.com/en/ documents/765514/the-rise-of-data-discovery-tools> [Accessed: 17 Dec. 2021].
29. A Closer Look at One of 2017's Most Important BI Trends. [online] Available at:<https://bi-survey.com/data-discovery> [Accessed: 17 Dec. 2021].
30. Magic Quadrant for Data Science and Machine-Learning Platforms / Peter Krensky, Erick Brethenoux, Carlie Idoine et al. [online] Available at:<https://www.gartner.com/ en/documents/3860063> [Accessed: 22 Feb. 2018].
31. Magic Quadrant for Data Science and Machine Learning Platforms. Carlie Idoine, Peter Krensky, Alexander Linden et al. [online] Available at:<https://www.gartner.com/en/documents/ 3899464/magic-quadrant-for-data-science-and-machine-learning-pla> [Accessed: 28 Jan. 2019].
32. Elliott, T., 2017. What is Artificial Intelligence Called?! [online] Available at: <https:// timoelliott.com/blog/2017/06/what-is-artificial-intelligence-called.html> [Accessed 11 Jul. 2019].
33. Sallam, R.L., Tapadinhas, J., Parenteau J. and et al. Magic Quadrant for Business Intelligence and Analytics Platforms. [online] Available at:<http://www.thgcfo.com/wp-content/uploads/2014/02/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms.pdf> [Accessed: 17 Feb. 2014].
34. Oursatyev, A.A., 2017. "*Big data*. Analytical databases and data warehouse: Teradata", Upravlausie sistemy i masiny, 2, pp. 51-67. (In Russian).
35. Harris, J. Five stages of data preparation. [online] Available at:<https://www.sas.com/ru_ua/insights/articles/ data-management/the-five-d-s-of-data-preparation.html> [Accessed: 17 Dec. 2021].
36. Ehtisham, Z., Sallam, R.L., Shubhangi, V. Market Guide for Data Data Preparation. [online] Available at:<https://www.gartner.com/doc/reprints?id=1-4FSMSCI&ct=170929&st=sb> [Accessed: 17 Dec. 2017].
37. Howard, Ph. Data Preparation (self-service). [online] Available at:<https://www.bloorresearch.com/technology/data-preparation-self-service/> [Accessed: 17 Dec. 2021].
38. BI Trends: Table of Contents. [online] Available at:<https://bi-survey.com/self-service-bi> [Accessed: 17 Dec. 2021].
39. The Definitive Guide to Self-Service Data. https://resources.boomi.com/resources/ home/the-definitive-guide-to-self-service-data> [Accessed: 17 Dec. 2021].
40. Oursatyev A.A., 2019. "*Big data*. Analytical databases and data warehouse: Greenplum", Upravlausie sistemy i masiny, 2, pp. 40 69. (In Russian).
41. Patel, M. Chorus Brings Data Science Minds Together. Feb., 2013. [online]. Available at: <https://blog.dellemc.com/en-us/chorus_data_science/> [Accessed: 17 Dec. 2021].
42. Greenplum Software Introduces Greenplum Chorus. Originally published April 12 2010. [online]. Available at: <http://www.b-eye-network.com/view/13182> [Accessed: 17 Dec. 2021].
43. Howard, Ph. Self-service data preparation and cataloguing. [online]. Available at: <https://www.bloorresearch.com/research/self-service-data-preparation-cataloguing-p2/> [Accessed: 12 Nov. 2016].
44. Howard, Ph. Data Preparation (self-service). [online]. Available at: <https: //www.bloorresearch.com/technology/data-preparation-self-service/> [Accessed: 01 July 2018].
45. Oursatyev, A.A., 2018. "*Big data*. Analytical databases and data warehouse: Vertica, Kdb". Upravlausie sistemy i masiny, 1, pp. 57-70. (In Russian).
46. Data profiling. [online]. Available at: <https://en.wikipedia.org/wiki/Data_profiling> [Accessed: 17 Dec. 2021].
47. Oursatyev, A.A., 2016. "SomeFrameworks forAnalytics Big Data". Upravlausie sistemy i masiny, 3, pp. 29 42. (In Russian).
48. Estensen, F.O. Master Data Management BI MicrosoftMDM. [online]. Available at: <https://ru.scribd.com/ presentation/252578258/BI-MicrosoftMDM-Frank-Olav-Estensen#scribd> [Accessed: 17 Dec. 2021].
49. Garter Glossary. Bimoda. [online]. Available at: <https://www.gartner.com/en/information-technology/glossary/bi-modal> [Accessed: 17 Dec. 2021].
50. Zeichick, A. Mode 1, Mode 2: Alan Zeichick on Bimodal Development. [online]. Available at: <https://blog.parasoft.com/mode-1-mode-2-alan-zeichick-on-bimodal-development> [Accessed: 17 Dec. 2015].

Received 06.07.2022

O.A. Oursatyev, Ph.D. Eng. Science, Senior Research Associate, International Research
and Training Centre for Information Technologies and Systems of the NAS and MES of Ukraine,
ORCID: https://orcid.org/ 0009-0009-8323-0525,
Acad. Glushkov ave., 40, Kiev, 03187, Ukraine,
aleksei@irtc.org.ua

DATA RESEARCH IN INDUSTRIAL DATA MINING PROJECTS IN THE BIG DATA GENERATION ERA

Introduction. The review material is based mainly on business intelligence (BI) solutions designed for the tasks with corporate data. But all the main aspects of working with data discussed in this research are also used on data processing platforms (Data Science Platform). Many BI vendors have expanded the capabilities of their systems to perform more advanced analytics, including Data Science. They added the phrase "Data Science" to their marketing research, and the term "advanced analytics" lost some popularity in relation to corporate data. The Data Science Platform provides a comprehensive set of tools for advanced users who traditionally work with data.

Capabilities that allow you to connect to multi-structured data across different types of storage platforms, both on-premises and in the cloud, and the infrastructure architecture of a modern BI analytics platform enable high-performance workloads, including business intelligence. It uses distributed architecture, massively parallel processing, data virtualization, in-memory computing, etc. The combination of traditional relational data processing with calculations on the well-known Apache Hadoop software infrastructure, which integrates a number of components of the Hadoop ecosystem (Apache Hive, HBase, Spark, Solr, etc.) with the necessary target functions. It allows you to create a fully functional platform for storing and processing structured and non-structures data.

Purpose. A review of data processing problems and an analysis of the use of world-class mathematical apparatus and tools for obtaining knowledge from information have been carried out.

Methods. The paper describes the use of Data Mining methods in big data processing tasks, as well as methods of business, recommendation and predictive analytics.

Result. The study suggests that machine learning-enhanced master data management (MDM), data quality, data preparation, and data catalogues will converge into a single, modern Enterprise Information Management (EIM) platform applicable to newest analytics projects. The analysis of the identifying useful data process can be beneficial to researchers and developers of the modern platforms for processing and researching data in various spheres of society.

Conclusion. A review of data processing problems and an analysis of the use of world-class mathematical apparatus and tools for obtaining knowledge from information have been carried out. It is shown that a high-quality solution to the problems of working with first-level data indicated in this review will be provided by data research in modern analytical platforms. Successful penetration into their essence at the level of obtaining knowledge using machine learning and artificial intelligence algorithms will make it possible to predict future results in managed objects (processes) and make informed decisions.

Keywords: *analytics maturity levels, predictive analytics, data discovery, data science, data preparation, self-service data, self-service data preparation platforms, data preparation conveyor.*