Іванов Дмитро,
доктор технічних наук, доцент,
професор кафедри комп'ютерних наук та інформаційних технологій,
Усата Олена,
кандидат педагогічних наук, доцент,
завідувач кафедри комп'ютерних наук та інформаційних технологій,
Житомирський державний університет імені Івана Франка,
м. Житомир, Україна

ВЕЛИКІ ДАНІ ТА ОСНОВИ РОБОТИ З НИМИ

В даний час обсяг даних, що генеруються машинами та людськими взаємодіями, швидко зростає, і технології розвиваються, намагаючись вирішити цю проблему. Глобальний ринок аналітики великих даних стрімко зростає з прогнозованою вартістю 745,15 мільярдів доларів США до 2030 року. Хоча великі дані широко обговорюються теоретично, існує ряд труднощів при їх обробці. Кращі дослідники та аналітики даних відрізняються визнанням різних типів аналітики, які найкраще відповідають потребам компанії. Відповідно метою статті є розглянути поняття великих даних та основні типи їх аналізу.

Словосполучення «big data» з'явилося в 2008 році у спецвипуску журналу Nature, де було опубліковано статті та думки щодо одного з найстрашніших викликів, що стоять перед сучасною наукою: великі потоки даних, які зараз генеруються. До поняття «big data» Кліффорд Лінч відніс будь-які масиви неоднорідних даних понад 150 Гб на добу. [1]

Зі статистичних звітів аналітичних агентств у 2010 році обсяги великих даних зазнали значного росту, і у 2012 році ці показники збільшились до 1,8 ЗБ (зеттабайт), а у 2015 – до 7 ЗБ і продовжують збільшуватись. Відповідно можна помітити, що з початку 2012-го обсяги даних зросли до величезних масштабів, і виникла потреба у їх систематизації та практичному застосуванні, тому Big Data опинились у центрі уваги провідних світових закладів вищої освіти, де навчають прикладним інженерним та IT-спеціальностям, й наукових установ. Потім значну увагу почали приділяти IT-корпорації, наприклад, такі як: Microsoft, IBM, Oracle, EMC, Google, Apple, Facebook та Amazon. Сьогодні великі дані використовують великі компанії у всіх галузях.

Щоб масив інформації позначити приставкою «big» він повинен мати такі ознаки, які компанія Meta Group запропонувала як основні характеристики

великих даних (Правило VVV) [2]:

- обсяг даних (Volume) – дані вимірюються за фізичною величиною та займаним простором на цифровому носії від 150 Гб на добу;
- швидкість накопичення та обробки масивів даних (Velocity) – інформація регулярно оновлюється і для обробки великих даних в реальному часі необхідні інтелектуальні технології;
- різноманітність типів даних (Variety) – інформація в масивах може бути не в однорідних форматах, частково або повністю структурованою або зберігатися безсистемно.

У сучасних системах розглядаються ще три додаткові фактори:

- мінливість (Variability) – потоки даних можуть мати піки та спади, сезонність, періодичність й відповідно сплески неструктурованої інформації потребують потужних технологій обробки;
- достовірність (Veracity) – достовірність як набору даних, так і результатів його аналізу;
- значимість даних (Value) – інформація може мати різну складність для сприйняття і обробки, що ускладнює роботу інтелектуальних систем (завданням стає визначення ступеня важливості вхідної інформації для швидкого структурування).

Принцип технології big data полягає в тому, щоб надати користувачам якомога більше інформації про будь-який предмет або явище. Завдання такого аналізу даних - проаналізувати всі переваги та недоліки, щоб прийняти правильне рішення. В інтелектуальних машинах на основі різних типів інформації будуються моделі майбутнього, моделюються різні варіанти та відстежуються результати. Джерелами даних виступають інтернет-блоги, соцмережі, сайти, ЗМІ та різноманітні форуми, транзакції, бази даних; показання метеорологічних приладів, супутників, датчики стільникового зв'язку, інтернет речей (IoT) та підключені до нього пристрої, статистика міст і держав про переміщення, народжуваність та смертність, медичні дані.

Сучасні обчислювальні системи забезпечують миттєвий доступ до масивів великих даних. Для їх зберігання використовують спеціальні дата-центри із найпотужнішими серверами, хмарні сховища, data lake – сховища великого обсягу неструктурованих даних з одного джерела тощо.

Для роботи з Big Data застосовують передові методи інтеграції та управління та підготовки даних для аналітики. Принципи роботи з масивами даних включають три основні фактори:

- розширюваність системи, під якою зазвичай розуміють горизонтальну масштабованість носіїв інформації, тобто у результаті зростання обсягів вхідних даних збільшується потужність та кількість серверів для їх зберігання;
- стійкість до відмови. Оскільки кількість цифрових носіїв та інтелектуальних машин може нескінченно збільшуватися пропорційно до обсягу даних, одним із чинників стабільної роботи з великими даними є стійкість до відмови серверів.
- локалізація. Окремі масиви інформації зберігаються та обробляються в

межах одного виділеного сервера, щоб заощаджувати час, ресурси, витрати на передачу даних.

Аналіз останніх доробків науковців та практиків показав, що наразі виділяють чотири основні типи аналітики Big Data [3] :

1. Описова аналітика (descriptive analytics). Ця аналітика найпоширеніша й аналізує як дані в реальному часі, так і історичні дані. Основна мета – знайти причини та закономірності успіху чи невдач у певній галузі та використати ці дані для побудови найбільш ефективних моделей. Описова аналітика використовує базові математичні функції. Типовими прикладами можна вважати соціологічні опитування та дані вебстатистики, які компанії отримують через Google Analytics.

2. Прогнозна або предикативна аналітика (predictive analytics) – допомагає спрогнозувати найбільш ймовірний розвиток подій на основі наявних даних. Використовуються готові шаблони, засновані на об'єктах або явищах зі схожими характеристиками. За допомогою цієї аналітики можна, наприклад, прорахувати обвал чи зміну цін на фондовому ринку, або оцінити можливості потенційного позичальника із виплати кредиту.

3. Приписна аналітика (prescriptive analytics). За допомогою великих даних і сучасних технологій можна виявити проблемні місця в бізнесі та інших видах діяльності і прорахувати, які сценарії допоможуть уникнути їх у майбутньому. Наприклад, за рахунок такої аналітики медичні центри можуть знизити кількість повторних госпіталізацій.

4. Діагностична аналітика (diagnostic analytics) – використовує дані, щоб проаналізувати причини того, що сталося. Це допомагає виявляти аномалії та випадкові зв'язки між подіями та діями. Наприклад, компанія з продажів може аналізувати дані про продажі та валовий прибуток для різних продуктів, щоб з'ясувати, чому вони принесли менше доходу, ніж очікувалося.

Також варто зазначити, що дані обробляють та аналізують за допомогою різних інструментів та технологій, які будуть детально розглядатись у подальшому.

Отже, чотири типи аналізу даних, описовий, діагностичний, прогнозний і приписний аналіз, є взаємопов’язаними рішеннями, які дають можливість компаніям оптимізувати свої великі дані. Кожен із цих типів аналітичних моделей надає унікальну перспективу, покращуючи операційні можливості організації.

Список використаних джерел та літератури

1. Nature. Sciense In Petabyte Era. Big Data. V.455. № 7209. 2008. URL: https://www.nature.com/nature/volumes/455/issues/7209.

2. Technology Industry 4.0. Big Data. URL: https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye

3. Four Types of Analytics with Example and Applications. ULR : https://www.projectpro.io/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209