

УДК 004.652

Богатирьова Т.О.¹, Зеленцова І.Я.², Паромова Т.О.³

¹студ. гр. КНТ-517 НУ "Запорізька політехніка"

²канд. техн. наук, доц. НУ "Запорізька політехніка"

³старш. викл. НУ "Запорізька політехніка"

ОГЛЯД ЗАСОБІВ РОБОТИ З BIG DATA

У зв'язку з розвитком інформаційних технологій, світові обсяги накопиченої інформації значно зростають з кожним роком. Так, в 2018 році ця цифра становила цілих 33 зеттатерабайта і, за прогнозами, до 2025 року людство сформує 175 36 інформації, говориться в доповіді The Data Age 2025, яка була підготовлена аналітиками компанії IDC в листопаді 2018 року [1].

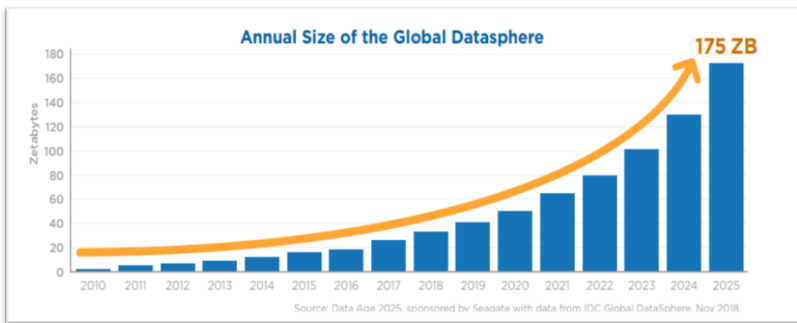


Рисунок 1 – Гістограма збільшення обсягів накопичених даних

Кількість користувачів глобальної мережі Інтернет значно зростає з кожним роком. Так, зокрема в звітах We Are Social і Hootsuite наводяться дані про глобальний стан цифрових технологій на 2019 рік: сьогодні в світі 5,11 мільярда унікальних мобільних користувачів, що на 100 мільйонів (2%) більше, ніж в попередньому році, в 2019 році аудиторія Інтернету налічує 4,39 мільярда чоловік, що на 366 мільйонів (9%) більше, ніж в січні 2018 року.

Це призводить практично до вибухового зростання обсягу накопиченої інформації, тому проблема обробки великих даних залишатиметься актуальною завжди і, як наслідок, призведе до значних якісних змін технологій обробки інформації. Як приклади джерел виникнення великих даних, зокрема наводяться дані, які безперервно надходять з вимірювальних пристроїв, події від радіочастотних ідентифікаторів, потоки повідомлень із

соціальних мереж, метеорологічні дані, дані дистанційного зондування Землі, потоки даних про місцезнаходження абонентів мереж стільникового зв'язку, пристроїв аудіо- і відеореєстрації. Очікується, що розвиток і початок широкого використання цих джерел ініціює проникнення технологій великих даних як в науково-дослідну діяльність, так і в комерційний сектор і сферу державного управління [2].

Big Data визначають як величезні обсяги неоднорідної цифрової інформації, що швидко надходить, які неможливо обробити традиційними інструментами [2]. Традиційні підходи до роботи з даними перестають бути ефективними в зв'язку з багатогранністю виникаючих завдань [3].

Визначальними характеристиками для великих даних, пов'язаних з особливостями їх обробки і аналізу, є принципи трьох, а потім розширених до п'яти і восьми V: volume, velocity, variety, veracity, viability, value, variability, visualization (фізичний обсяг, швидкість приросту даних і необхідності їх швидкої обробки, можливість одночасно обробляти дані різних типів, достовірність, життєздатність, цінність, мінливість, візуалізація відповідно).

На відміну від традиційних підходів, основні принципи роботи з великими даними такі:

- горизонтальна масштабованість;
- відмовостійкість;
- локальність даних.

Багато з того, що стосується великих даних, підходів, які використовуються для їх аналізу, насправді існує досить давно. Наприклад, обробка зображень з камер спостереження, коли аналізується не одна картинка, а потік даних.

Міжнародна консалтингова компанія McKinsey, що спеціалізується на вирішенні завдань, пов'язаних зі стратегічним управлінням, виділяє 11 методів і технік аналізу, які можна застосувати до великих даними:

- data Mining;
- краудсорсінг;
- змішання та інтеграція даних;
- машинне навчання;
- штучні нейронні мережі;
- розпізнавання образів;
- прогнозна аналітика;
- імітаційне моделювання;
- просторовий аналіз;
- статистичний аналіз;
- візуалізація аналітичних даних.

Необхідність порівняно недорогого зберігання і обробки гігантських

обсягів неструктурованої інформації привела до створення спеціалізованого програмного забезпечення, яке дозволило розподіляти дані по кластерам з сотень і тисяч вузлів, а також обробляти їх в паралельному режимі. Це призвело до створення Hadoop - відкритого фреймворка фірми Apache Software Foundation, який дозволяв створювати розподілені системи на базі відносно недорогого обладнання масового попиту. На його основі фірми - розробники Cloudera, MapR і Hortonworks пропонують свій дистрибутив. У 2011 році Hadoop помітили Oracle, IBM, Teradata і також включили в лінійки своїх програмних продуктів.

Для звичної роботи з інформацією за допомогою SQL на базі Hadoop розгортаються СУБД Hive або Impala. При потребі в можливостях NoSQL рішення - HBase.

Зберігати інформацію також можна в хмарі Amazon S3 або в NoSQL базі даних Cassandra, управляти ресурсами кластерів за допомогою Apache Mesos, видобувати і обробляти дані за допомогою Apache Spark, який здатний працювати і поза Hadoop. Spark останнім часом набирає популярність, так як обіцяє прискорення виконання розподілених програм обробки даних в порівнянні з Hadoop MapReduce до 100 разів. Може працювати як поверх HDFS під керуванням Hadoop YARN, так і поза фреймворка Hadoop на базі Cassandra, Amazon S3 і Google Cloud Storage (через Alluxio).

Дані, отримані з розподілених систем, обробляються в аналітичних інструментах SAS (Enterprise Miner), IBM (SPSS), Teradata (Aster Analytics) або Oracle (Advanced Analytics) або за допомогою маси інших комерційних та open-source рішень.

Big Data допомагають вирішувати різні завдання в різних сферах людського життя, серед них: ритейл, медицина, фінанси, промисловість, енергетика, туризм, екологія, розваги. Завдяки обробці та аналізу великого масиву даних, поліпшується якість товарів і послуг, з'являються нові можливості для розвитку бізнесу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. The Digitization of the World From Edge to Core [Електронний ресурс]. – Режим доступу: <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf>
2. Big data [Електронний ресурс]. – Режим доступу: <https://intellect.icu/big-data-6821>
3. Big Data [Електронний ресурс]. – Режим доступу: <https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye>