

Бурятів Олексій Олексійович, здобувач вищої освіти КПІ ім. Ігоря Сікорського, Україна
Науковий керівник: Вітківська Ірина Іванівна, старший викладач кафедри інформатики та програмної інженерії, КПІ ім. Ігоря Сікорського, Україна

ПОБУДОВА ІНФРАСТРУКТУРИ ДЛЯ ВИСОКОПРОДУКТИВНИХ СИСТЕМ ТА ВПРОВАДЖЕННЯ ТЕНДЕНЦІЙ З ВИКОРИСТАННЯ BIG DATA ІНСТРУМЕНТІВ

INFRASTRUCTURE DEVELOPMENT FOR HIGH-LOAD SYSTEMS AND IMPLEMENTATION TRENDS OF USING BIG DATA TOOLS

Анотація. Діджиталізація процесів у повсякденному житті людини призводить до зростання цифрових носіїв інформації якими вона користується. Це призводить до необхідності обробки даних, які генерується у великих об'ємах щосекундно. Для обробки надвеликих масивів даних необхідні високі потужності, які зуміють виконати поставлену задачу. Попередні інструменти, якими користувалися інженери, стають неефективними в порівнянні з сьогоденними тенденціями. Нові методи та підходи допоможуть справитися з будь-якою задачею та знизити ризики втрати даних або недоступність сервісів. Інструменти по впровадженню та підтримці програмного забезпечення різноманітний, і особливо важливо правильно визначити доречність використання в тій чи іншій задачі.

КЛЮЧОВІ СЛОВА: Big Data, високонавантажені системи, автоматизація, Kubernetes, відмовостійкість.

Abstract. Digitalization of processes in everyday human's life leads to increasing of usage smart devices, which they use. It leads to necessity of data processing, which generates in large amounts every second. In order to process enormous arrays of data, it is necessary to have high power, which will be able to perform the task. Previous tools, which used by engineers, become ineffective compared to current trends. New methods and approaches will help to cope with any task and reduce the risk of data loss or unavailability of services. Tools for implementing and maintaining software are diverse and it is especially important to correctly determine the appropriateness of use in a particular tasks.

KEY WORDS: Big Data, high-load systems, automation, Kubernetes, high availability.

Вступ. За статистикою, Big Data є однією з найбільш зростаючих галузей інформаційних технологій. Обсяг зберігання та отримання даних, згідно зі статистикою, збільшується в два рази кожні 1,2 роки. Наприклад, між 2012 та 2014 роками, обсяг даних, що щомісяця передаються мобільними мережами, збільшився на 81% [1, с. 195]. У 2014 році, обсяг мобільного трафіку склав 2,5 ексабайта на місяць, а вже у 2019 році - 24,3 ексабайта, за даними Cisco. Застосування технологій Big Data дуже широке. За результатами опитування IBM Institute, 53% використання Big Data стосується сфери клієнтського

сервісу, 40% - операційної ефективності, а 7% - управління ризиками.

Таким чином, Big Data – це прогресуюча сфера технологій, яка набула поширення у багатьох сферах бізнесу і відіграє важливу роль у розвитку компаній.

Основна частина. Big Data (з. англ. – великі дані) – надвеликі масиви даних, які обробляються спеціальними програмними та технічними засобами, а також потребують високопродуктивного обладнання для швидкого та безперебійного аналізу даних [2, с. 12]. Зазвичай Big Data рішення потребують великі бізнеси, які збирають статистику зі

своїх продуктів, щоб у подальшому покращити сервіси надання послуг, використовуючи аналітику.

Розглядаючи можливі варіанти побудови потужного сховища та системи збору й обробки інформації, один із основних принципів – це створення потужного кластеру. На сьогоднішній день є купа рішень програмного забезпечення для Big Data продуктів – як безкоштовних, так і комерційних. Якщо обирати із open-source – це створення кластеру Kubernetes, Apache NiFi, Apache Spark, Confluent Kafka. Якщо обирати із комерційних пропозицій – це Cloudera Data Platform, Teradata тощо. Кожен сервіс розроблявся під окремі задачі, тому побудова універсального кластеру неможлива.

Однією із найпопулярніших платформ для управління ПЗ – це Kubernetes [3, с.1]. Головними перевагами використання сервісу є:

1. High availability – високий поріг відмовостійкості у разі виходу з ладу фізичного серверу, пошкодження обладнання, недоступність мережі тощо.

2. Open source – можливість використовувати безкоштовно, що є вагомим чинником для сучасного бізнесу, а також постійне вдосконалення користувачами цього сервісу, які зацікавлені в покращенні та усуненні багів;

3. Ізоляція ресурсів – кожне ПЗ запускається в окремому «віртуальному» середовищі, яке нічого не знає про інші застосунки;

4. Можливість розгортання в хмарі – багато провайдерів хмарних послуг (такі як Microsoft Azure, Amazon Web Services, Google Cloud Platform) інтегрували Kubernetes до свого сховища і тому є можливість використання цього сервісу в хмарі.

З чого починати при створенні високопродуктивних систем для обчислення надвеликих масивів даних? Перш за все необхідно вирішити питання з наявними ресурсами, а саме: обрати фізичні сервери (або хмарного провайдера сервісів) та їх потужність, визначити пропускну здатність

мережевого каналу, підрахувати об'єми та швидкість дискового простору.

Автоматизація – це необхідна частина розробки, впровадження, оновлення та інтегрування застосунків і сервісів. Працюючи в команді, необхідно максимально автоматизувати процеси, адже кожен учасник розробки повинен розуміти як працює ціла система та можливість легко й швидко працювати разом. Для цього існують декілька інструментів, які допоможуть автоматизувати процеси від написання коду, зборки образів та розгортання застосунків, а також для створення самої інфраструктури як на наземному кластері, так і в хмарі:

1. Terraform – програмне забезпечення, яке використовується для створення, редагування, видалення та версіонування хмарної інфраструктури. Цей інструмент передбачає більше ніж 200 хмарних провайдерів та наземної інфраструктури [4, с. 5];

2. Ansible – інструмент по автоматизації виконання коду для великих кластерів фізичних серверів. За допомогою скриптів, або як ще називають – playbooks (з англ. – сценарій) можна одночасно виконувати поставлені задачі, зберігаючи час та консистентність інфраструктури [4, с. 6];

3. CI/CD – (continuous integration and continuous deployment – з англ. – неперервна інтеграція та неперервне розгортання) – практика, яка автоматизує процес тестування та розгортання програмного забезпечення під час написання коду. Цей інструмент важливий тим, що він реагує на зміни в коді, написані розробником, а також має логування й версіонування, адже при неуспішних тестах інфраструктура зберігає свій попередній працюючий стан. Такі інструменти є в GitLab, GitHub, Jenkins, TeamCity тощо.

Якщо розглядати вищеперераховані інструменти у площині високопродуктивних систем та обчислень, то вони просто необхідні – в інакшому випадку є ймовірність недоступності застосунків, що може вкрай негативно вплинути на роботу усіх налаштованих процесів, адже часто один процес залежить від іншого.

Висновки. Отже, підсумовуючи вищесказане, можна прийти до висновку, що Big Data – це один із головних напрямів сучасного бізнесу, який має великі масиви даних для обробки. Для створення потужних обчислень великих масивів даних, необхідно потурбуватися про високонавантажену систему та сервіси, які будуть займатися обробкою. Одним із основних інструментів є Kubernetes, який оркеструє запущені сервіси, ізолюючи один від одного, контролює використання ресурсів під кожен процес та гарантує високу доступність навіть при збоїв системи. Усім цим необхідно керувати інструментами автоматизації, які пришвидшують планові та позапланові оновлення, логують дії та роблять їх консистентними.

Список інформаційних джерел

1. Глоба Л.С., Курдеча В.В., Дундяк Р.Р. , РОЗВАНТАЖЕННЯ МОБІЛЬНИХ МЕРЕЖ ЗА ДОПОМОГОЮ ПОЄДНАННЯ ТЕХНОЛОГІЙ WI-FI OFFLOAD ТА LTE. ХЕНДОВЕР МІЖ ДВОМА ТЕХНОЛОГІЯМИ [Електронний ресурс] / Національний технічний університет України «Київський політехнічний інститут» – Режим доступу до ресурсу: <http://molodyvcheny.in.ua/files/journal/2016/6/48.pdf>
2. В. Г. Саріогло. “Великі дані” як джерело інформації та інструментарій для офіційної статистики: потенціал, проблеми, перспективи [Електронний ресурс] / В. Г. Саріогло // ISSN 2519-1853 СТАТИСТИКА УКРАЇНИ. – 2016. – Режим доступу до ресурсу: http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21REF=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=su_2016_4_5
3. Eunsook Kim; Kyungwoon Lee; Chuck Yoo. On the Resource Management of Kubernetes [Електронний ресурс] / Eunsook Kim; Kyungwoon Lee; Chuck Yoo // IEEE. – 2021. – Режим доступу до ресурсу: <https://ieeexplore.ieee.org/abstract/document/9333977>.
4. Single and Multi-Cloud Disaster Recovery Management using Terraform and Ansible [Електронний ресурс] // National College of Ireland Project Submission Sheet School of Computing. – 2019. – Режим доступу до ресурсу: <https://norma.ncirl.ie/4144/1/rambaraththiyagarajan.pdf>