

РОЗГОРТАННЯ ВЛАСНИХ АРТЕФАКТІВ ЕКОСИСТЕМИ HADOOP

UDC 004.45+004.62

N. Melnyk; A. Lutskev, Ph.D., Associate Professor

DEPLOYMENT OF OWN ARTIFACTS OF THE HADOOP ECOSYSTEM

Ключові слова: Великі Дані, Hadoop, артефакт, Bigtop, Ambari.

Key words: Big Data, Hadoop, artifact, Bigtop, Ambari.

Зростання обсягу даних у світі робить критично важливим використання програмного забезпечення для їх ефективної обробки та управління [1]. Однією з найпоширеніших платформ для роботи з Big Data є екосистема Apache Hadoop, яка забезпечує можливості для обробки великих обсягів структурованих і неструктурованих даних [2]. Одним із важливих аспектів у цьому контексті є розгортання власних артефактів Hadoop, що дозволяє адаптувати платформу під конкретні потреби організацій. Артефакти Hadoop – це сформовані пакети програмних компонентів екосистеми Hadoop.

Актуальність дослідження методу розгортання власних артефактів екосистеми опрацювання великих даних Hadoop обумовлена необхідністю забезпечення ефективної та надійної роботи з даними. Системи, побудовані на основі Hadoop, використовуються в різних сферах, таких як аналітика структурованих і неструктурованих даних, машинне навчання та управління ресурсами.

Основною метою використання методу розгортання власних артефактів Hadoop виступає забезпечення більшої незалежності від зовнішніх корпорацій та розробників, що надають свої рішення, а також забезпечення можливості повного налаштування екосистеми під власні потреби компанії методом перетворення і переконфігурування наявних рішень для власних потреб або створення власних сервісів і інтегрування їх у інфраструктуру.

Використання таких інструментів, як Apache Bigtop для побудови та тестування артефактів, та Apache Ambari для управління кластерами, дає змогу значно підвищити продуктивність шляхом переконфігурування сервісів платформи з урахуванням власних вимог та забезпечити високу якість розгорнутих систем.

Apache Bigtop виступає важливим інструментом для побудови й тестування артефактів екосистеми Hadoop. Він забезпечує можливість пакування та перевірки сумісності компонентів, що гарантує стабільність і ефективність роботи розгорнутих систем [3].

Для формування артефактів Apache Bigtop спочатку клонує необхідний GitHub репозиторій сервісу і застосовує спеціальні git-патчі для внесення необхідних змін у локальний репозиторій, не змінюючи при цьому оригінальні origin-гілки на GitHub. Після цього він запускає необхідну команду для формування проєкту і пакує двійкові файли у інсталяційні пакети (deb або rpm).

Існує три основних варіанти переналаштування екосистеми Hadoop з використанням Apache Bigtop:

- створення власних відгалужень або використання відгалужень інших організацій GitHub репозиторіїв оригінальних сервісів та внесення змін безпосередньо в код програм, після чого необхідно змінити посилання на репозиторій у файлі налаштування “bigtop.bom”;
- створення власних git-патчів для внесення змін у локальний репозиторій, клонований з оригінального GitHub репозиторію;
- реалізація та інтеграція власних сервісів у стек. Даний метод передбачає повну розробку власного сервісу (створення нового проєкту, його структури, написання програмних скриптів, цифрове підписування файлів тощо), а також написання тестів та реалізацію механізму пакування двійкових файлів.

Одним із ключових завдань при роботі з екосистемою Hadoop є розгортання та управління кластером. Apache Ambari спрощує цей процес, надаючи інструменти для налаштування, моніторингу та повного управління кластером Hadoop. Ambari забезпечує зручний інтерфейс для встановлення, конфігурації та оновлення сервісів Hadoop.

Для розгортання артефактів Ambari використовує попередньо визначені стеки. Stack Ambari – це сукупність компонентів та технологій, які використовуються самим програмним рішенням Ambari для спрощення та автоматизації розгортання, налаштування, управління та моніторингу окремих сервісів Hadoop та у загальному кластері Hadoop [4]. Кожен компонент стеку виконує певну роль при розгортанні та взаємодії з іншими компонентами для досягнення спільної мети. Стеки можуть мати власні версії, а також вони забезпечують власні інструменти для керування версіями сервісів. Кожна версія стеку пропонує попередньо визначений перелік сервісів певних версій для забезпечення сумісності. Варто окремо відзначити, що для забезпечення сумісності необхідно використовувати вже наявні API та протоколи, реалізовувати обробку даних у форматах, визначених та підтримуваних сервісами із якими ведеться взаємодія, забезпечувати безпеку тощо. Для перевірки сумісності необхідно проводити тестування та перевіряти роботу як кожного з сервісів, так і взаємодію між ними перед використанням у кінцевому робочому середовищі (Production).

Використання таких інструментів, як Apache Bigtop та Apache Ambari, є критично важливим для побудови і розгортання артефактів екосистеми Hadoop. Завдяки цим інструментам можна досягти високого рівня продуктивності та стабільності роботи систем, що опрацьовують великі обсяги даних. Власні артефакти дають змогу адаптувати Hadoop до конкретних завдань організації, забезпечуючи при цьому ефективну обробку і зберігання даних.

Література

1. Джавад А. Ш., Мухаммад А. Х. Big Data Systems. A 360-degree Approach. Лондон: Taylor & Francis, 2023. 340 с.
3. Apache Hadoop. Apache Software Foundation. URL: <https://hadoop.apache.org/>, (дата звернення: 20.11.2024).
4. Apache Bigtop. Apache Software Foundation. URL: <https://bigtop.apache.org/>, (дата звернення: 22.11.2024).
5. Defining a Custom Stack and Services. Confluence Apache Ambari. URL: <https://cwiki.apache.org/confluence/display/AMBARI/Defining+a+Custom+Stack+and+Services>, (дата звернення: 04.12.2024).