УДК 004.45+004.62
Мельник Н. – ст. гр. СІм-52
Тернопільський національний технічний університет імені Івана Пулюя

МЕТОДИ ТА ЗАСОБИ РОЗГОРТАННЯ АРТЕФАКТІВ ЕКОСИСТЕМИ HADOOP

Науковий керівник: доцент, кандидат технічних наук Луцків А. М.

Melnyk N.
Ternopil Ivan Puluj National Technical University

METHODS AND MEANS OF DEPLOYING ARTIFACTS OF THE HADOOP ECOSYSTEM

Supervisor: Associate Professor, PhD A. Lutskiv

Ключові слова: кластер, Hadoop, артефакт, великі дані
Keywords: cluster, Hadoop, artifact, Big Data

Every day the amount of different types of data produced by humanity increases greatly. Therefore, software tools for data management and processing are of great importance. Due to its volume, velocity and variety, this data can be treated as a Big Data [1]. Nowadays, appropriate software tools mostly belong to Apache Hadoop ecosystem [2] which comprises a lot of tools and libraries developed by different software engineering teams with open source licenses. To provide reliable, resilient and efficient platform for data processing these tools should be consistent in the following aspects: library versions, configuration and deployment. The main objective of this research is to provide an approach to resolve this task.

Big Data management is cyclical in nature. It can be generally divided into five phases: capture, organize, integrate, analyze, and act (Fig. 1). Hadoop components are specially designed to perform tasks on each of the listed phases [3].

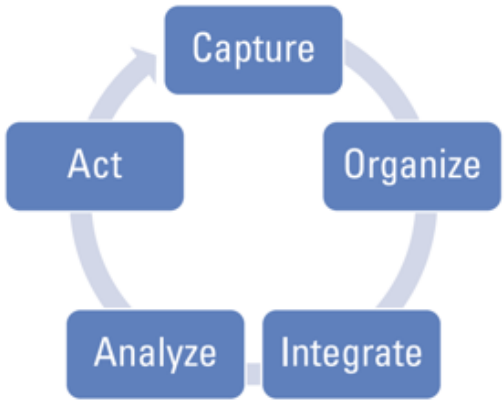

Figure 1. Cycle of Big Data Management [2]

The Hadoop ecosystem is a collection of freely distributed open-source software products designed for processing and analyzing large volumes of data. This ecosystem emerged in response to the need for efficient handling and storing large volumes of structured

and unstructured data, which is constantly growing. The main components of the Hadoop ecosystem are Hadoop Distributed File System (HDFS), MapReduce, Hadoop YARN (Yet Another Resource Negotiator), and many additional projects, that extend the capabilities of Hadoop core components for various use cases. The most popular of these additional projects are Spark, Zookeeper, Hive, HBase, and Kafka.

Hadoop artifacts are built packages of software components of the Hadoop ecosystem. Efficient deployment and testing of Hadoop artifacts are critically important for several reasons, such as system reliability, performance optimization, resource utilization, data security, and core quality assurance.

It is also worth noting separately the Apache Ambari and Apache Bigtop projects, which are related to research on the given topic.

The main goal of Ambari [4] is to provide provisioning, management, monitoring, and operating Hadoop clusters at scale. Ambari gives a central location to examine the status, configuration, and resource usage for each one of these. For provisioning a Hadoop cluster, Ambari provides a step-by-step wizard for installing Hadoop services across any number of hosts. Ambari handles the configuration of Hadoop services for the cluster. For managing the cluster, Ambari provides central management for starting, stopping, and reconfiguring Hadoop services across the entire cluster. For monitoring a Hadoop cluster, Ambari provides a dashboard for monitoring the health and status of the Hadoop cluster, leverages a metric system for collecting metrics, and also leverages an alert framework to notify the administrators when nodes go down, disk space is low, or anything requires an administrator's attention.

Apache Bigtop [5] serves as a comprehensive platform for packaging, testing, and deploying Hadoop ecosystem components. It provides a uniform build and testing environment for various Hadoop projects, ensuring compatibility and interoperability across the ecosystem. Also, this project allows customization of Hadoop distributions to suit specific requirements.

In summary, Apache Ambari and Apache Bigtop play pivotal roles in simplifying cluster management, ensuring quality assurance, fostering collaboration, and driving innovation within the Hadoop ecosystem. Their adoption is essential for organizations looking to harness the full potential of Hadoop for Big Data processing and analytics. The result of this research will be a fully configured and tested cluster with Apache Hadoop ecosystem, taking into account performance and data security issues as well.

References

1. Jawwad A. S., Muhammad A. K. Big Data Systems. A 360-degree Approach. Chapman & Hall, 2021.

2. Apache Hadoop. Apache Software Foundation. URL: https://hadoop.apache.org/, accessed 1 May 2024.

3. Kaur I. Cycle of Big Data Management. (2020) URL: https://ishmeetk10.medium.com/cycle-of-big-data-management-ac9b99899a44.

4. Apache Ambari. Apache Software Foundation. URL: https://ambari.apache.org/, accessed 1 May 2024.

5. Apache Bigtop. Apache Software Foundation. URL: https://bigtop.apache.org/, accessed 1 May 2024.