УДК 003.26.09; 004.032.24-004.272.3

А.М. Луцків, канд.техн.наук, доц., В.М. Діденко
Тернопільський національний технічний університет імені Івана Пулюя, Україна

КЛЮЧОВІ ОСОБЛИВОСТІ ПЛАТФОРМИ HADOOP 3

A.M. Lutskiv Ph.D., Assoc. Prof., V.M. Didenko

KEY FEATURES OF HADOOP 3 PLATFORM

При розв'язанні задач зі сфери Big Data (великих даних) стандартом де-факто вважається використання компонентів екосистеми Hadoop. Хоча ці компоненти є окремими програмними системами, належать різним розробницьким проектам та можуть бути використані окремо, здебільшого використовуються у сукупності. Такий підхід до використання зумовлений необхідністю побудови цілісних та багатофункціональних систем, які дають змогу будувати складні процеси опрацювання різнотипних даних: отримання, попереднє опрацювання (фільтрацію, відображення одних типів в інші тощо), опрацювання/трансформацію (об'єднання з іншими даними, обчислення тощо), аналітичне опрацювання (статистика, машинне навчання), формування результату, візуалізація. Це опрацювання може відбуватись у пакетному режимі (з часовими затримки), в потоковому режимі або гібридно.

Найбільші інтегратори Hadoop — Cloudera та Hortonworks, пропонують готові дистрибутиви програм, які поєднують відповідні компоненти: Cloudera (CDH 6.x) та Hortonworks (Hortonworks Data Platform 3 та Hortonworks Data flow 3)[1]. Використання готових дистрибутивів суттєво спрощує роботу інженера й розгортання багатовузлового кластера може зайняти, лише, кілька годин (залежно від особливостей конфігурації та швидкості використовуваного обладнання).

Розглянемо переваги, які надає Hadoop 3 у порівнянні з попередньою версією:

- Hadoop 3 повністю базується на Java 8, що надає набагато ширші можливості розробникам при використанні API.
- Ефективніший алгоритм зберігання резервних копій даних забезпечує зменшення надлишкового простору з 200% до 50%. Зберігається не повна копія резервних даних, а виключно дані, які необхідні для відновлення у випадку аварійних ситуацій.
- Вдосконалене масштабування служби планувальника завдань YARN Timeline Service, що дає змогу будувати масштабовані кластерні системи з більшою кількістю вузлів ніж у Hadoop 2.
- Спрощені засоби адміністрування кластером, які були забезпечені шляхом рефакторингу Shell-сценаріїв запуску сервісів кластера.
- Ізоляція CLASSPATH клієнта при використанні різних бібліотек на стороні кластера та клієнта.
- Оптимізація MapReduce на 30% у задачах передавання даних між Map- та Reduce-виконавцями шляхом написання платформозалежного коду.
- Підвищення надійності роботи кластера шляхом можливості використання більше двох NameNode-вузлів.
- Підтримка хмарних сховищ даних Microsoft Azure Data Lake та Aliyun Object Storage System.
- Балансування дискового навантаження не лише на міжвузловому рівні усього кластера, а й у середині кожного вузла (DataNode) при використанні кількох дисків.
- Спрощення процесу конфігурування Heap-пам'яті для MapReduce-задач.
- Підтримка GPU-ресурсів у планувальнику ресурсів YARN.

Платформа Hadoop 3 забезпечує сумісність з іншими компонентами (Таблиця 1).

Таблиця 1 - Сумісність програмних систем з платформою Hadoop 3

Apache Project	Version	Призначення компоненту
HBase	2.0.0	NoSQL розподілене сховище зберігання даних
Spark	2.0	Обчислювальний фреймворк для задач аналізу даних, машинного навчання, потокового опрацювання даних
Hive	2.1.0	Система пакетного опрацювання даних, яка надає SQL-доступ до даних, що зберігаються в HDFS
Oozie	5.0	Планувальник задач для Hadoop
Pig	0.16	Система пакетного опрацювання даних, яка надає доступ до даних, що зберігаються в HDFS за допомогою мови Pig Latin
Solr	6.x	Пошуковий фреймворк, який надає можливості повнотекстового пошуку
Kafka	0.10	Розподілена система опрацювання повідомлення з великою пропускною здатністю

Таким чином до ключових особливостей компонентів платформи Hadoop належать:

- можливість реалізувати на їх основі лямбда- або каппа-архітектури[2];
- можливість інтеграції з більшістю хмарних сервісів шляхом використання різноманітних програмних інтерфейсів;
- широка підтримка провайдерами хмарних сервісів (Amazon Web Services, Google Cloud Platform, IBM Cloud Data Services, Oracle BigData Cloud Services, Microsoft Azure та іншими);
- можливість побудови програм для опрацювання великих обсягів даних на мові Java та сучасних фреймворків (Spring, чи інших JEE-сумісних), що дає змогу відносно просто інтегрувати інструментарій аналітики великих даних у Enterprise-вирішення;
- використання доступних апаратно-програмних засобів[3].

Література

1. Hortonworks Data Platform 3.0.1. Release Notes 3. [Електронний ресурс] Режим доступу: URL: https://docs.hortonworks.com/HDPDocuments/HDP3/HDP-3.0.1/release-notes/hdp-release-notes.pdf

2. Samizadeh I. A brief introduction to two data processing architectures—Lambda and Kappa for Big Data / Iman Samizadeh// Medium. Towards Data Science. [Електронний ресурс] Режим доступу: URL: https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb

3. Загородна Н. В., Лупенко С. А. Луцків А. М. Обґрунтування вибору доступних програмно-апаратних засобів високопродуктивних обчислювальних систем для задач криптоаналізу. // Електроніка та системи управління. 2011. №1(27). - К.: НАУ, 2011. - с.42-50.