Міністерство освіти і науки України

Черкаський державний технологічний університет

Кафедра програмного забезпечення автоматизованих систем

**ЗВІТ**

з лабораторної роботи №1

з предмету «Аналіз великих даних»

|  |  |
| --- | --- |
| Перевірив:  к. т. н., д. каф. ПЗАС  Рідкокаша А.А.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 р. | Виконав:  студент 1-го курсу  групи МПЗ-1904  Гаврилюк В. Є. |

Черкаси 2019

**Лабораторна робота №1**

**Тема:** Вивчення засобів фреймворку — Hadoop MapReduce, Hadoop, Spark.

**Мета:** Ознайомитися з реалізацією технології обробки великих даних Hadoop за програмною моделлю MapReduce.

**Завдання:**

1. Форматувати вузол кластера.
2. Запустити службу управління кластером.
3. Запустити службу управління файловою системою.
4. Перевірити запуск і переглянути інформацію про вузол (node).

**Короткі теоретичні відомості**

Apache Hadoop — вільна програмна платформа і каркас для організації розподіленого зберігання і обробки наборів великих даних з використанням моделі програмування MapReduce, при якій завдання ділиться на багато дрібніших відособлених фрагментів, кожен з яких може бути запущений на окремому вузлі кластера, що складається з серійних комп'ютерів. Всі модулі в Hadoop спроектовані з врахуванням припущення, що апаратне забезпечення часто виходить з ладу і такі ситуації повинні автоматично опрацьовуватись фреймворком.

Ядро системи Apache Hadoop складається з розподіленої файлової системи Hadoop Distributed Filesystem (HDFS), та системи обчислень на основі моделі програмування MapReduce. Hadoop розділяє файли на великі блоки і розподіляє їх між вузлами кластера. Тоді він передає запакований код на вузли для паралельної обробки даних. Цей підхід користується локальністю даних, коли вузли маніпулюють лише даними до яких мають доступ. Це дозволяє обробляти набір даних швидше і ефективніше ніж в традиційнішій суперкомп'ютерній архітектурі яка покладається на паралельну файлову систему в якій обчислення та дані для них передаються через високошвидкісну мережу.

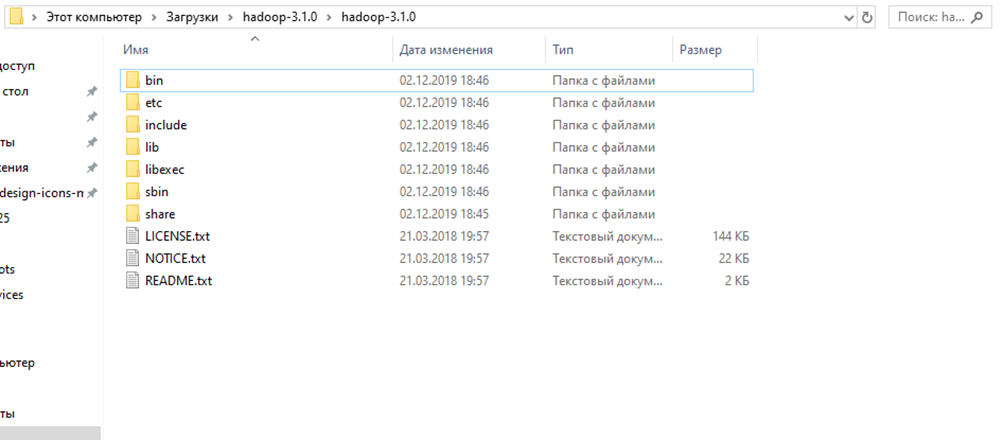
Основний фреймворк Apache Hadoop складається з наступних модулів:

* Hadoop Common — містить бібліотеки та утиліти потрібні іншим модулям Hadoop;
* Hadoop Distributed File System (HDFS) — розподілена файлова система, яка зберігає дані на звичайних машинах, надаючи дуже високу загальну пропускну здатність на кластері загалом;
* Hadoop YARN — платформа що відповідає за керування обчислювальними ресурсами в кластерах і їх використання для користувацьких завдань;
* Hadoop MapReduce — реалізація моделі програмування MapReduce для обробки великих об'ємів даних.

З часом, термін Hadoop почав вживатись не тільки щодо вищезгаданих базових модулів та підмодулів, а й до «екосистеми», тобто набору додаткових пакетів програмного забезпечення, які можуть встановлюватись поверх, або поряд з Hadoop, наприклад таких як Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache ZooKeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache Oozie, та Apache Storm. MapReduce та HDFS в Apache Hadoop's були натхненними статтями Google про їх алгоритм MapReduce та Google File System. Фреймворк Hadoop написаний переважно на Java, з частиною системного коду на C та утилітами командного рядка як shell скрипти. Хоча в програмах MapReduce звичайним є код на Java, для реалізації «map» та «reduce» частин користувацької програми можна використовувати будь-яку мову програмування завдяки «Hadoop Streaming». Інші проекти в екосистемі Hadoop надають багатші інтерфейси користувача.

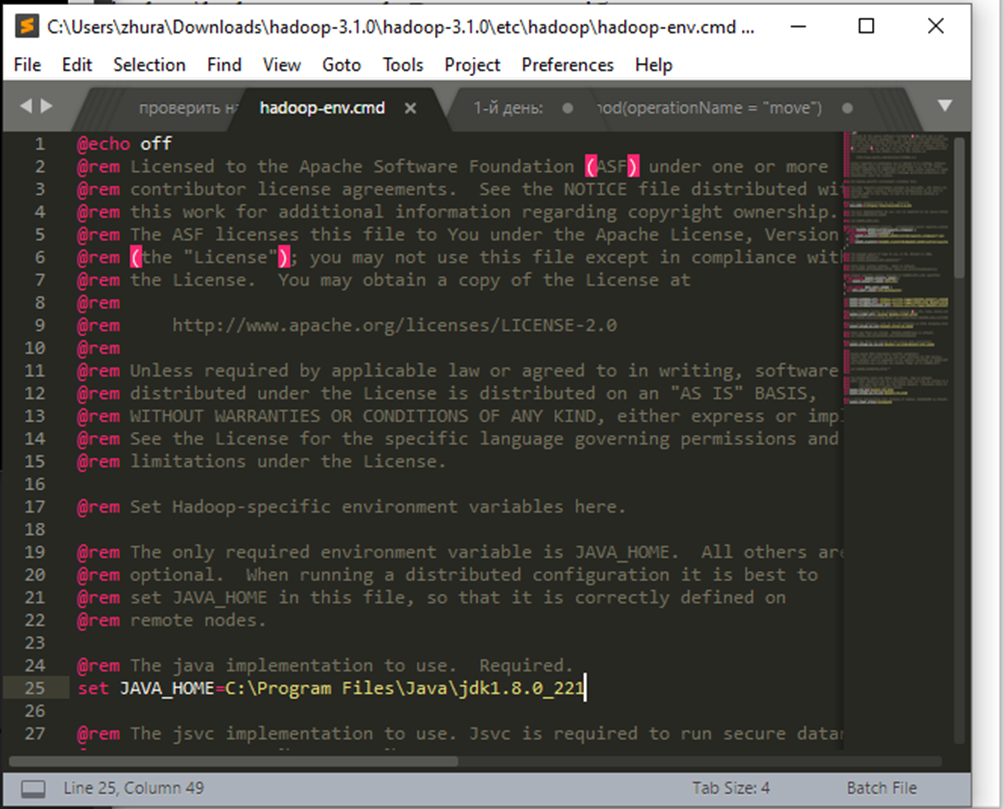
**Хід роботи**

1. Завантажуємо Hadoop для Windows 7/8/10. Для встановлення необхідно розпакувати архів (рис. 1).

****

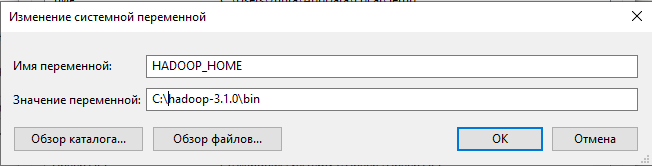
*Рис. 1 – Директорія з файлами Hadoop.*

1. При відсутності на комп'ютері Java, необхідно скачати та встановити.
2. Відкриваємо файл у редакторі: C:\Users\[username]\Downloads\hadoop-3.1.0\etc\hadoop\hadoop-env.cmd (рис. 2). В ньому потрібно прописати шлях до бібліотек віртуальної машини Java.

****

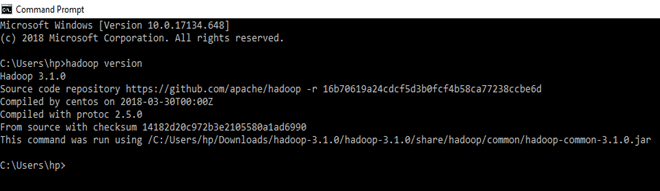
*Рис. 2 – Вміст файлу hadoop-env.cmd.*

1. Також потрібно прописати змінну в розділ системних змінних (рис. 3).



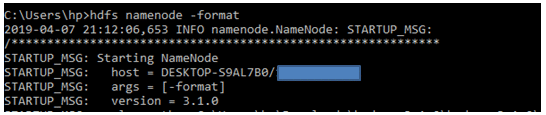
*Рис. 3 – Запис змінної в системні.*

1. Для перевірки версії Hadoop потрібно ввести в консолі команду hadoop version (рис. 4).

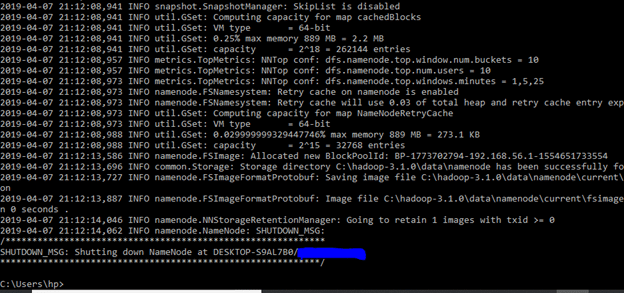


*Рис. 4 – Перевірка версії Hadoop.*

1. Для форматування NameNode потрібно ввести команду hdfs namenode –format (рис. 5 та 6).

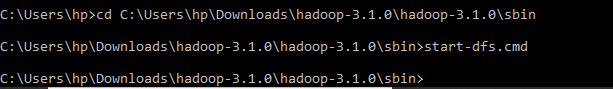


*Рис. 5 – Старт форматування NameNode.*

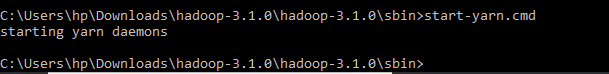


*Рис. 6 – Процес форматування NameNode та його завершення.*

1. Далі потрібно змінити дерикторію в консолі (рис. 7). Потім потрібно запустити наступну команду (рис. 8).

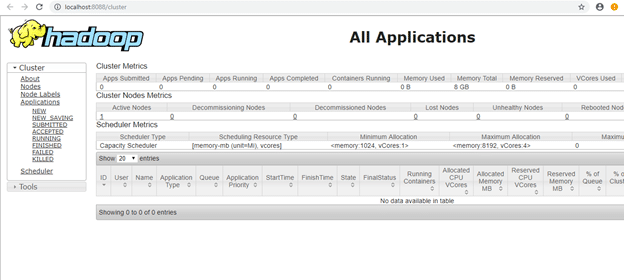


*Рис. 7 – Перехід в іншу директорію.*



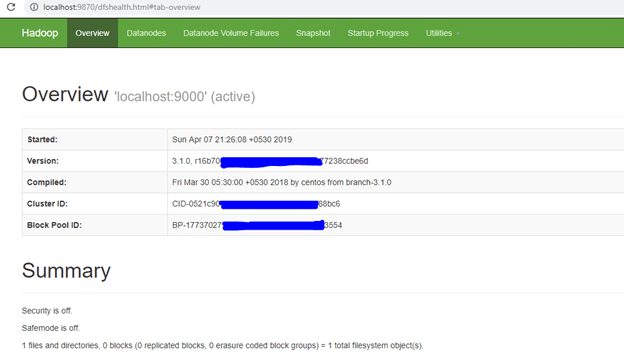
*Рис. 8 – Запуск yarn команди.*

1. Щоб подивитися інформацію про ресурси, поточні роботи і роботи, які завершилися з помилками, потрібно перейти за посиланням в браузері - http://localhost:8088/cluster (рис. 9).



*Рис. 9 – Виведення інформації про інформацію про ресурсита роботи.*

1. Щоб перевірити інформацію про hdfs (namenode & datanode) потрібно перейти за посиланням в браузері - http://localhost:9870/ (рис. 10).



*Рис. 10 – Перевірка інформації про hdfs.*

**Висновок:** Під час виконання лабораторної роботи я ознайомився з реалізацією технології обробки великих даних Hadoop. Також встановив та провів попередні налаштування системи Hadoop. Дізнався про призначення даної системи та її використання.