УДК 004.4

Бугайов Олександр Сергійович
студент
Національного технічного університету України
«Київський політехнічний інститут імені Ігоря Сікорського»
Бугаев Александр Сергеевич
студент
Национального технического университета Украины
«Киевский политехнический институт имени Игоря Сикорского»
Buhaiov Oleksandr
student of the
National technical university of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

ОГЛЯД ЕКОСИСТЕМИ HADOOP

ОБЗОР ЭКОСИСТЕМЫ HADOOP

HADOOP ECOSYSTEM REWIEW

Анотація. *Розглянута програмна платформа для роботи з великими обсягами даних Hadoop та технології для інтеграції з нею, що використовуються для вирішення різноманітних дослідницьких задач.*

Ключові слова: *Big Data, Hadoop, дані.*

Аннотация. *Рассмотрена программная платформа для работы с большими объемами данных Hadoop и технологии для интеграции с ней, которые используются для решения различных исследовательских задач.*

Ключевые слова: *Big Data, Hadoop, данные.*

Summary. *Program platform for work with Big Data – Hadoop is reviewed with technologies for integration with it, which can be applied for resolving different research tasks.*

Key words: *Big Data, Hadoop, data.*

Вступ. Оскільки різноманітність і обсяг даних, накопичених компаніями, продовжує зростати значними темпами, то також зростає і популярність Hadoop. Це обумовлено тим, що ця платформа дає змогу зберігати та обробляти величезні обсяги неструктурованих даних через не надто вартісне обладнання.

Основи Hadoop. Існує дві першочергові речі, які потрібно знати про Hadoop. По-перше, Розподілена файлова система Hadoop (HDFS) дозволяє зберігати файли надвеликого об'єму — таблиці з мільярдами записів, що розміщуються на десятках (а в деяких випадках — тисячах) дешевих серверах. До того ж, він може бути використаний безліч різноманітних файлів таким самим чином. Це пояснює, чому HDFS використовують компанії, що працюють з найбільшими у світі наборами даних — Facebook, Google, IBM. По-друге, парадигма MapReduce — це алгоритм Hadoop обробки та аналізу великих об'ємів даних, якими керує HDFS. MapReduce перевертає традиційний принцип аналізу даних. Замість того щоб збирати дані з десятків чи сотень серверів та передавати їх через мережу, MapReduce переміщає програмне забезпечення до даних, виконуючи певні обчислення паралельно. Ці два компонента разом зробили Hadoop одним з найпоширеніших інструментів для BigData, що використовують як великі компанії так і стартапи. Популярність Hadoop обумовлена наявність великої кількості ПЗ, яке допомагає значно розширити базові можливості Hadoop, та утворює цілу екосистему навколо нього.

Вдосконалення MapReduce. Базова парадигма MapReduce — потужний інструмент для аналізу вели-

ких обсягів даних, але він має певні недоліки, найголовніший з яких — це інструмент низького рівня, тому він вимагає написання великої кількості коду для виконання стандартних задач. Це обумовило створення деяких мов обробки даних, що компілюються у MapReduce. Серед них:

- Hive — розробка Facebook, створений з метою додати до неструктурованого Hadoop SQL-подібні можливості. Головною перевагою є те, що він дозволяє розробникам виконувати ad-hoc запити без знання того, як влаштований MapReduce. Hive написаний на HiveQL, що базується на SQL. Таким чином, якщо набір даних містить багато структурованих табличних даних — Hive буде доречним.
- Pig — розробка Yahoo, інструмент для інженерів, що потребують глибокого аналізу та контролю для їх процесів обробки даних. Ця платформа використовує процедурну мову, Pig Latin, що дозволяє інженерам визначати потік даних на кожному кроці. Звідси одна з переваг Pig — простота відлагодження. Також, оскільки Pig базується на мові більш високого рівня, вона включає набір вбудованих функцій, що дозволяють інженерам керувати даними та проводити базовий аналіз, не потребуючи написання програм MapReduce.
- Crunch — бібліотека для Java, створена Apache, дозволяє розробникам зі знаннями Java використовувати потужні та ефективні інструменти для написання застосунків з MapReduce.

NoSQL можливості. Стандартні засобі Hadoop підходять для оффлайн та пакетної обробки «холодних» даних (тих, що не використовуються, історичних). При роботі з даними, доступ до яких відбувається в момент роботи, виникає потреба у більш продуктивних інструментах. Для цього використовується Cassandra — розподілена NoSQL база даних. При цьому для роботи з цією БД використовується структурована мова схожа на SQL. В результаті отримана система має високу продуктивність та майже постійні показники доступності даних, що робить її системою для мобільних та веб-застосунків, що використовують опитування в реальному часі. Інша значна перевага Cassandra — майже лінійне масштабування. Зі збільшенням обсягів даних додаються нові вузли до кластера.

Аналітика та Машинне навчання. Для організацій, що навантажені великим обсягом даних та потребують проведення просунутого аналізу над цими даними, також є кілька інструментів, що інтегруються разом з Hadoop:

- Milb — бібліотека Spark, що має підтримку таких операцій як кластеризація, регресія, класифікація, спільна фільтрація та розмірнісна редукція.
- Mahout — бібліотека алгоритмів для машинного навчання, що розроблена спеціально для роботи над Hadoop. Це добре задокументована бібліотека, що дозволяє інженеру швидко та ефективно аналізувати дані та знаходити паттерни.

Збір та розміщення даних. Hadoop може використовуватись також з великими обсягами даних, що постійно збільшуються. Для того, щоб скерувати потік нових даних до сховища можна використовувати спеціальні інструменти:

- Flume — використовує систему «агентів» щоб зібрати та скерувати дані в HDFS або іншу систему, що використовується. Ці агенти налаштовуються на очікування певних подій, які вони захоплять на направлять у відповідні канали для запису в сховище.
- Sqoop — інструмент що створений для роботи саме з Hadoop, який переміщує дані між HDFS та реляційними базами даних.

Висновок. Таким чином, Hadoop — це комплексне рішення для роботи з великими обсягами даних, функціонал якого може бути розширений з використанням наявних для інтеграції технологій.

Література

1. Keenan T. Get to Know the Hadoop Ecosystem [Електронний ресурс] / Tyler Keenan / UpWork — Режим доступу до ресурсу: https://www.upwork.com/hiring/data/get-to-know-hadoop-ecosystem/.

2. Who uses Hadoop [Електронний ресурс] / Apache — Режим доступу до ресурсу: https://wiki.apache.org/hadoop/PoweredBy.

3. Apache Hadoop [Електронний ресурс] / Wikipedia. — 2017. — Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Apache_Hadoop.