Щадин Евгений Евгеньевич
магистрант
Научный руководитель
Панферова Елена Викторовна
канд. техн. наук, доцент

ФГБОУ ВО «Тульский государственный
педагогический университет им. Л.Н. Толстого»
г. Тула, Тульская область

DOI 10.31483/r-111123

HADOOP ДЛЯ ХРАНЕНИЯ И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Аннотация*: в статье рассматривается технология Hadoop как инструмент для хранения и обработки больших объемов данных. Описываются основные принципы работы Hadoop, его архитектура и компоненты. Представлены преимущества использования Hadoop для анализа и обработки больших данных, а также его применение в современных информационных системах.*

Ключевые слова*: Hadoop, Big Data, хранение объемов данных, обработка объемов данных, информационная система.*

На рубеже XX–XXI веков произошел значимый рост использования цифровой информации. Размеры файлов увеличиваются, информационные технологии, такие как социальные сети, средства связи, электронный документооборот, безналичный расчет, а также системы видеонаблюдения, большие данные (Big Data), становятся неотъемлемой частью нашей жизни. Это, очевидно, приводит к увеличению ресурсов на технологии хранения и обработки данных, что требует активных исследований и разработок в этой области со стороны научного сообщества [1].

«Большие данные – это сбор и анализ большого набора данных, который содержит множество интеллектуальных и необработанных данных, основанных

на пользовательских данных, показаниях датчиков, медицинских и корпоративных данных» [2].

Проблемы при работе с Big Data.

1. Большой объем данных требует дорогостоящих технологий для их хранения и обработки.

2. Защита данных для обеспечения их целостности, доступности и конфиденциальности.

3. Неупорядоченный формат хранения данных, где каждый элемент имеет уникальный вид.

4. Сложность структуризации, сортировки и поиска элементов в общей системе.

5. Огромная скорость поступления данных, как следствие, скорость их обработки, значительно уступающая скорости поступления, что может привести к долгому ожиданию ответов и устареванию информации в процессе обработки.

Все эти факторы вызвали необходимость специализированного программного обеспечения для обработки и анализа огромных объемов данных, с которыми традиционные системы не могли справиться. Значительный рост информации, требующей параллельной обработки, появление новых типов данных и желание снизить стоимость их обработки послужили основной мотивацией для создания Hadoop.

Этот программная платформа, набор инструментов, облегчающий разработку, структурирование и объединение различных компонент масштабного программного проекта (framework), предоставляет масштабируемую распределенную инфраструктуру для работы с Big Data, обеспечивая возможность обработки данных на кластере узлов и снижая затраты на обработку информации.

Hadoop как проект фонда Apache Software Foundation был официально представлен 1 апреля 2006 году и продолжает активно развиваться. В 2008 году был выпущен первый стабильный релиз Hadoop, и с тех пор он постоянно обновляется и улучшается.

Приведем хронологию ключевых моментов в развитии Hadoop:

– 2005: проект Hadoop в первые был реализован как часть поискового движка Nutch, разрабатываемого Doug Cutting и Mike Cafarella;

– 2006: Yahoo начала использовать Hadoop для своих поисковых и рекламных систем;

– 2008: Apache Hadoop стал открытым проектом Apache Software Foundation;

– 2009: появление Hadoop 0.20. В этот момент многие компании, включая Facebook, начали активно использовать Hadoop;

– 2012: Hadoop 1.0.0 заявлен в виде версии stable. Это выступило ключевым моментом в развитии фреймворка;

– 2013: представлен Hadoop 2.0 (также известный как YARN), добавивший новые возможности для работы с данными и архитектурой фреймворка;

– 2017: на ряду с последующими релизами Hadoop 2.x развивались новые проекты в этой системе. Так же в этот период развивались новые проекты в экосистеме Hadoop, такие как Apache Spark и Apache Flink;

– 2018: анонсирован как available первые стабильные релизы Apache Hadoop 3.x.

– 2024: в настоящее время особенностями доступные на настоящее время версии 3.4.0, являются [3]:

1) HDFS RBF: поддержка хранилища токенов на основе RDBMS;

2) «федерация» на основе маршрутизатора HDFS теперь поддерживает хранение токенов делегирования в MySQL, HADOOP-18535, что улучшает работу токенов по сравнению с исходной реализацией на основе Zookeeper;

3) новые API файловой системы: HADOOP-18671 перенес ряд API-интерфейсов, специфичных для HDFS, в Hadoop Common, чтобы обеспечить возможность запуска определенных приложений, зависящих от семантики HDFS, в других файловых системах, совместимых с Hadoop. В частности, функции RecoveryLease и isFileClosed доступны через интерфейс LeaseRecoverable, а setSafeMode – через интерфейс SafeMode.

Hadoop предлагает богатый функционал для обработки, хранения и анализа больших данных.

1. Hadoop Distributed File System (HDFS) – распределенная файловая система, разработанная для обработки больших объемов данных в распределенной среде. Она предоставляет масштабируемое хранилище для данных, которые могут быть разбиты на блоки и распределены по разным узлам в кластере.

2. MapReduce: MapReduce – программная модель и инфраструктура для обработки и генерации больших наборов структурированных и неструктурированных данных. Она позволяет распределять задачи обработки данных по узлам кластера и эффективно обрабатывать данные параллельно.

3. Yet Another Resource Negotiator (YARN) – универсальная распределенная система управления ресурсами, которая управляет выполнением задач в кластере. Она расширяет возможности Hadoop, позволяя запускать не только MapReduce-задачи, но и другие типы приложений, такие как Spark, Flink и др.

4. Hadoop Ecosystem. Hadoop обладает обширной экосистемой, включающей различные проекты и инструменты для обработки и анализа данных, такие как Apache Hive (для выполнения SQL-запросов), Apache Pig (для написания данных потоков обработки), Apache Spark (для обработки данных в памяти).

5. Hadoop способен обрабатывать петабайты (1 Пбайт – единица измерения количества информации, равная 1015 (квадриллион) байт) данных на тысячах распределенных серверов без значительной потери производительности.

Hadoop используется для хранения и обработки данных.

1. Facebook использует Hadoop для обработки и анализа огромных объемов пользовательских данных. Это помогает улучшить персонализацию рекомендаций и оптимизировать алгоритмы новостной ленты.

2. eBay внедрил Hadoop для анализа клиентских предпочтений, улучшения работы поиска товаров и оптимизации рекомендательных систем.

3. Netflix использует Hadoop для анализа поведения зрителей, что позволяет выпускать более релевантный контент и персонализированные рекомендации.

4. Airbnb использует Hadoop для анализа данных бронирований, предпочтений пользователей и прогнозирования спроса на жилье.

Hadoop используется для хранения, организации и обработки информации на ресурсах, где требуется оперативная массово-параллельная обработка данных.

На данный момент у Hadoop есть несколько аналогов: MapR, Cloudera, DataProc. Сравнение систем представлено в таблице 1.

Таблица 1

Сравнение систем

	MapR	Cloudera	Apache Hadoop	DataProc
Тип расположения	На территории предприятия	На территории предприятия	На территории предприятия	В облаке или на территории предприятия
Сложность миграции	Нет миграции	Высокая	Высокая	Очень высокая
Риск	Нет	Нет после миграции	Нет после миграции	Низкая
Разнообразие экосистемы	Средняя	Большая	Большая / развивающаяся	Большая / развивающаяся
Гибкость	Низкая	Низкая	Низкая	Высокая
Коммерческая	Да	Да	Нет	Да

В ходе сравнения были разобраны основные преимущества и недостатки этих систем, Hadoop имеет ряд преимуществ над другими системами.

1. Имеет возможность миграции.

2. Имеет большой набор развивающихся фреймворков входящих в экосистему.

3. Проект свободно распространяемый.

Hadoop решает ряд проблем, вызванных особенностями Big Data, таких как распределенное хранение, параллельная обработка, отказоустойчивость и удобный доступ к данным. Hadoop позволяет предприятиям эффективно работать с данными, помогая принимать обоснованные решения. Преимущества Hadoop

включают гибкость, масштабируемость и открытый исходный код, что делает его одним из наиболее популярных и востребованных инструментов для работы с большими данными в настоящее время.

Список литературы

1. Менщиков А.А. Основные проблемы использования больших данных в современных информационных системах / А.А. Менщиков, В.Э. Перфильев, М.Ю. Федосенко [и др.] // Столыпинский вестник. – 2022. – Т. 4. №1 – EDN NDZPQO

2. Царев Ю.В. Создание и исследование характеристик работы распределенного кластера Hadoop / Ю.В. Царев, В.С. Качайло, А.Ю. Кокорина // Вестник науки. – 2022. – №6 (51) [Электронный ресурс]. – Режим доступа: https://clck.ru/3AkigY (дата обращения: 19.03.2024).

3. Apache Hadoop [Электронный ресурс]. – Режим доступа: https://hadoop.apache.org/ (дата обращения: 18.03.2024).