

ОБЗОР ИНФОРМАЦИОННЫХ СИСТЕМ ПО СБОРУ И ОБРАБОТКЕ БОЛЬШИХ ДАННЫХ HADOOP И STORM**Захарец Ольга Ивановна**

студент, Волгоградский государственный социально-педагогический университет, РФ, г. Волгоград

Маньшин Максим Евгеньевич

научный руководитель, канд. пед. наук, доцент, Волгоградский государственный социально-педагогический университет, РФ, г. Волгоград

Аннотация. В работе рассматриваются технологии сбора и обработки больших данных, их преимущества и недостатки. Перспективы их развития.

Ключевые слова: Big Data; технология HADOOP; технология STORM.

В настоящее время активно увеличивается объём информации, которая генерируется при помощи различных компьютерных систем, это можно увидеть даже на примере обычных камер видеонаблюдения, у которых поток данных может достигать нескольких десятков петабайт в секунду. Традиционные подходы к обработке столь крупного объема данных не справляются с анализом такого интенсивного потока информации из-за ограничений на размер хранилищ данных и задержки на сохранение информации перед обработкой. В связи с этой проблемой появились технологии, позволяющие организовать работу по сбору и обработке Big Data.

Первым, кто ввел понятие Big Data в сфере информационных технологий, был редактор журнала Nature Клиффорд Линч в спецвыпуске 2008 года. В своей статье редактор отметил, что видит в дальнейшем большой рост объема информации в мире. К Big Data Линч отнес любые массивы неоднородных данных более 150 Гб в сутки, однако единого критерия до сих пор не существует.

Big Data -это группа технологий и методов производительной обработки динамически растущих объемов данных в информационных технологиях [1];

На сегодняшний день существует множество различных технологий по сбору и обработке Big Data, но все они обладают рядом определенных особенностей [5]: работа с информацией огромного объема и разнообразного состава; информация постоянно обновляется и размещается в разных источниках; возможность ведения аналитики для выявления практических знаний, которые можно монетизировать в прибыль; наглядное отображение отчетов и возможности сценарного анализа; цель применения технологий Big Data - увеличение эффективности работы, создание новых продуктов и повышение конкурентоспособности.

Именно такими особенностями обладают технологии HADOOP и STORM. Рассмотрим каждую из этих технологий более подробно.

Hadoop – это программное обеспечение с открытым исходным кодом, которое используется для обработки Big Data. Данная технология имеет широкое применение среди организаций для анализа больших данных [2].

Впервые Hadoop появился как проект в 2006 под руководством Каттинга в корпорации Yahoo, команда которой занималась разработкой инфраструктуры распределённых вычислений, а уже в 2008 корпорация, под управлением Hadoop, запустила кластерную поисковую машину на 10 тысяч процессорных ядер [6].

Сегодня, несмотря на то, что сфера информационных технологий по сбору и обработке Big Data активно развивается и появляются новые и усовершенствованные системы, Hadoop, обладающая открытым кодом и управляемая Apache Software Foundation, всё-так же продолжает активно использоваться. Представленная технология по-прежнему служит инструментом для осуществления масштабируемых и надежных вычислений, но также он используется в качестве файлового хранилища [1]. Таким образом можно выявить ряд преимуществ данной технологии: снижение времени на обработку данных; снижение стоимости оборудования; повышение отказоустойчивости; линейная масштабируемость; работа с неструктурированными данными. Несмотря на все выше перечисленные преимущества, процесс развития не стоит на месте и у технологий появляются новые возможности, которыми уже не обладает Hadoop, в связи с этим можно выявить ряд недостатков использования данной технологии:

- Проблемы безопасности;
- Недостаток удобства для пользователя;
- Не подходит для аналитики в реальном времени.

Технология Storm, также обладает открытым исходным кодом, но главное её отличием от Hadoop в том, что она является распределительной системой вычислений в реальном времени, главная её роль заключается в надежной обработке неограниченных потоков данных. Технология Storm была создана компанией BackType, а затем после покупки в 2011 году перешла к Twitter* и была открыта как проект распределительных вычислений, под лицензией Eclipse Public License [3].

Поскольку эта система является строго потоковой и используется для обработки событий одно за другим оно имеет ряд преимуществ:

- Высокая пропускная способность;
- Низкая задержка;
- Интеграция с любыми системами управления очередью и брокерами сообщений, а также базами данных;
- Отказоустойчивость;
- Гарантия обработки данных;

Но, так же, как и любая система имеет свои недостатки:

- Отсутствует управление состоянием;
- Отсутствие возможностей гибкой обработки событий;
- Поддержка минимум однократной доставки сообщений (at least once).

Таким образом обе технологии имеют большой потенциал в современном высокотехнологичном мире, они обладают рядом преимуществ, опираясь на которые, каждый желающий может выбрать для своей работы с Big Data наиболее соответствующую систему. Так технология Hadoop чаще всего используется в поисковой и контекстной механизмах высоконагруженных веб-сайтов и интернет-магазинов, таких как Yahoo!, Facebook*, Google**, AliExpress, Ebay и др. А вот Storm лучше всего использовать, при обработке данных в реальном времени, как например это делают Twitter*, NaviSite и Wego.

*(социальные сети, запрещенные на территории РФ, как продукты организации Meta, признанной экстремистской – прим.ред.)

******(По требованию Роскомнадзора информируем, что иностранное лицо, владеющее информационными ресурсами Google, является нарушителем законодательства Российской Федерации" – прим.ред)

Список литературы:

1. Богданов А.В., Тхуреин К.Л., Пья С. К., Чжо З. Сравнение производительности инструментов для обработки больших данных.- М.: Современные наукоемкие технологии.- 2020.- URL: <https://www.elibrary.ru/item.asp?id=43030035> (дата обращения: 05.05.22)
2. Дзидзава Э.Т., Ахмедов К.М. Большие данные и Hadoop: обзорный доклад.- М.: Вестник магистратуры.- 2021.- URL: <https://cyberleninka.ru/article/n/bolshie-dannye-i-hadoop-obzornyy-doklad> (дата обращения 08.05.22)
3. Казанский Н.Л., Проценко В.И., Серафимович П.Г. Сравнение производительности систем потокового анализа данных в задаче обработки изображений скользящим окном.- М.: Компьютерная оптика.- 2014.- URL: <https://cyberleninka.ru/article/n/sravnenie-proizvoditelnosti-sistem-potokovogo-analiza-dannyh-v-zadache-obrabotki-izobrazheniy-skolzyaschim-oknom> (дата обращения 05.05.22)
4. Клиффорд Линч Science in the petabyte era.- М.: Nature.-2008.- URL: <https://www.nature.com/nature/volumes/455/issues/7209> (дата обращения 22.04.22 г.)