УДК 004.6 **Технические науки**

Федоренко Николай Сергеевич, *магистрант 1 курса*
Хакасский государственный университет им. Н.Ф. Катанова
Хрусталев Виталий Игоревич, *научный руководитель, канд. техн. наук*
Хакасский государственный университет им. Н.Ф. Катанова

БОЛЬШИЕ ДАННЫЕ. ПОДХОДЫ К ТОЛКОВАНИЮ ТЕРМИНА

Аннотация: в статье описывается проблема к подходу определения больших данных. Указывается на нефизическую природу термина. Описывается модель трех V. Дается некоторый ретроспективный анализ появления термина.

Ключевые слова: большие данные, анализ данных.

Abstract: the article describes the problem to the approach of determining large data. Indicates the nonphysical nature of the term. The model of three V. Some retrospective analysis of the appearance of the term is given.

Keywords: big data, data analysis.

«Большие данные», в течение уже более десяти лет, стало словосочетанием, которое употребляется во многих выступлениях, как на крупных IT конференциях, так и медийной фразой, которую можно услышать на различных экономических форумах. При этом отсутствует четкое толкование данному термину, как в целом в научном сообществе, так даже у тех, кто использует данную концепцию в своих целях. Не смотря на наличие в словосочетании термина «данные», отсутствует описание и классификация того, какие объемы информации будут относиться, а какие нет к данной категории.

Данный факт может показаться изначально странным, однако если проанализировать существующую ситуацию с постоянным ростом общего объема информации, то отсутствие привязки к конкретным объемам информации становится понятным.

Термин «большие данные» появился в работах Джона Маша [1] еще в конце 90-х годов. В целом почти все работы, посвященные проблеме больших объемов информации, сводились к вопросу обработки большого количества информации, и к явным ограничениям вычислительной техники по ее обработке. Эта проблема возникла задолго до общей компьютеризации и вообще появления компьютерной техники. Еще в 1944 г. Библиотекарь Уэслианского университета Фримонт Райдер опубликовал «Ученый и будущее научной библиотеки. Проблема и решение» [2]. В оценках Райдера число американских университетских библиотек должно удваиваться каждые 16 лет, таким образом, он предполагал, что к 2040 году в Йельской библиотеке будет располагаться около 200 миллионов томов, которые будут занимать более 6000 миль полок, требующих для каталогизации более 6000 тысяч человек персонала.

Как видно, в целом, проблемы с данными сводятся к тому, что их количество постоянно растет и возникает вопрос по их структурированию и обработке. В настоящее время компьютерная техника позволяет сравнительно быстро сохранять и обрабатывать терабайты информации, однако постоянно возникает вопрос о важности, нужности и полезности информации.

Наиболее удачным и точным на наш взгляд является определение больших данных, построенных на модели «трех V»: volume (объем), velocity (скорость) и variety (разнообразие) [3]. Объем информации, необходимый для отнесения ее к большим данным, разнится в понимании отдельных авторов, но стоит говорить о том, что мы переходим к концепции больших данных, когда речь идет об обработке отдельных датасетов, размер которых занимает гигабайты, терабайты и больше (данные сравнительно актуальны на дату написания текущего текста). Вторая V в модели означает скорость получения,

обработки, анализа и интерпретирования данных. В настоящее время нет никакого сомнения, что концепция больших данных предполагает работу в режиме близком к реальному времени. Разнообразие данных (третья V модели) указывает на разнородную структуру исходных данных, почти вся информация, получаемая в большом объеме, редко является хорошо структурированной, в конечном итоге исследователю или практику в области анализа данных приходится обрабатывать информацию с разной степенью структурированности.

Наличие в модели лишь одной характеристики (объем), отвечающей за количественное выражении информации, указывает на нефизическую природу термина «большие данные».

Подводя итог, стоит отметить, что большие данные нельзя выразить численно или обозначить нижнюю границу для отнесения данных к категории больших. Большие данные предполагают концепцию, в которой данные образуются с такой скоростью, что для их обработки необходимо применять высокоскоростные технологии, построенные на принципе распараллеливании задачи и решению отдельных ее частей на разных вычислительных устройствах.

Библиографический список:

1. John R. Mashey (25 April 1998). "Big Data ... and the Next Wave of InfraStress" (PDF). Slides from invited talk. Usenix. Retrieved 28 September 2016. [электронный ресурс] http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf (дата обращения 11.06.2018).

2. Fremont Rider The Scholar and the Future of the Research Library: A Problem and Its Solution Hadham Press, 1944, p. 236.

3. Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety [электронный ресурс] https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (дата обращения 11.06.2018).