## Лабораторная работа №1

Тема: Характеристика больших данных организации

Цель: изучение организации как источника и накопителя больших данных, приобретение навыков предварительного анализа больших данных организации на предмет их переработки с помощью технологий Big Data.

Теоретические положения

**Формула Big Data: 7V**

Суть технологий Big Data – это работа с гигантскими массивами данных (что следует, впрочем, уже из самого термина). Но объем данных еще не делает погоды. Аналитики придумали емкую формулу Big Data – они считают, что в определении проекта Big Data должны фигурировать семь важных характеристик, «7 V»: Volume, Velocity, Variety, Veracity, Variability, Visualization, Value. То есть объем, скорость, разнообразие, достоверность, изменчивость, визуализация, ценность. При этом каждая «V» важна для понимания общей картины.

Меньше всего вопросов вызывают первые три «V»: Volume, Velocity, Variety. Действительно, кто будет спорить, что Big Data – это прежде всего объем, Volume? Объем данных растет по экспоненте: например, самолеты ежегодно генерируют 2,5 млрд ТБ данных с датчиков, установленных в двигателях. При этом данные постоянно обновляются, генерятся новые, и скорость обновления (Velocity – вторая «V») также важна для того, чтобы считать их «большими». Например, каждую минуту в мире выполняется почти 2,5 миллиона запросов к поисковой системе Google. Задача проектов Big Data заключается в том, чтобы справиться с огромной скоростью, с которой данные создаются, и анализировать их в режиме реального времени.

Третья «V» – Variety, разнообразие. Это означает, что проекты Big Data должны включать данные в самых разных форматах: структурированные и неструктурированные данные, текстовые, графические, данные корпоративной почты или соцсетей, вплоть до видео. Каждый из этих типов данных требует различных типов анализа и подходящих инструментов. Социальные медиа могут помочь владельцам бренда проанализировать настроения клиентов, а сенсорные данные предоставят информацию о том, как чаще всего используется продукт, чтобы применить эти знания для его улучшения.

Еще недавно трех «V» было вполне достаточно. Но все на свете изменяется, в том числе и подходы к определению. Поэтому аналитики присовокупили еще четыре «V», чтобы избежать недопонимания. Итак, в определение были добавлены Veracity, Variability, Visualization, Value. Рассмотрим каждый из этих пунктов.

4. Veracity – Достоверность: безусловно, эта характеристика является крайне важной, поскольку любой анализ будет совершенно бесполезен, если данные окажутся недостоверными. Более того, нам крайне важно заранее удостовериться, что с данными все ок, ведь их неточность может привести к неправильным решениям. Самый простой пример – контакты с ложными именами и неточной контактной информацией.

5. Variability – Изменчивость: новое веяние в сфере Big Data. Здесь речь идет о том, что значение одних и тех же данных может различаться в зависимости от контекста, например, одни и те же слова в Твиттере могут иметь различные значения и отражать различные настроения. Мы должны учитывать все нюансы! Для того чтобы выполнить правильный анализ настроений, алгоритмы должны быть в состоянии понять контекст и быть в состоянии расшифровать точное значение слова в этом контексте.

6. Visualization – Визуализация: это необходимая часть анализа, поскольку именно визуализация делает большие данные доступными для человеческого восприятия. Визуализация больших объемов сложных данных гораздо более эффективна и понятна для человека, чем электронные таблицы и отчеты, полные чисел и формул. Конечно, визуализация в рамках Big Data не означает построение обычных графиков или круговых диаграмм: возможно, будут построены сложные графики, которые будут включать в себя множество переменных данных, однако они все равно останутся понятными и читаемыми.

7. Value – Ценность: здесь речь идет о том, чтобы извлечь максимум пользы из результатов анализа больших данных. Важно то, как вы будете использовать эти данные и сможете ли превратить свою организацию в продвинутую компанию, которая опирается на идеи, полученные из анализа данных, для принятия решений.

Задание

1. Дать краткую характеристику организации, выбранной в качестве объекта исследования.
2. Идентифицировать и охарактеризовать по формуле 7V большие данные, генерируемые, циркулирующие и накапливающиеся в организации. Выделить недостатки существующей технологии обработки данных.
3. Обосновать необходимость и целесообразность применения технологий распределенной обработки данных для переработки больших данных организации.

Требования к содержанию отчета.

В пункте 1 дается характеристика организации и ее подразделения(й), причастных к переработке больших данных. При этом не нужно приводить общие организационные структуры управления и давать им описание. Приводится краткая информация о структуре организации и решаемых задачах.

В пункте 2 с помощью предлагаемых теоретических материалов и рекомендуемых материалов сети Интернет приводятся правила организации данных, претендующих на статус больших. Дается характеристика уровню управления этими данными в контексте формулы 7V. При этом следует отметить, какие из приведенных в пункте 1 задач решаются недостаточно эффективно в связи с недостаточным качеством переработки выделенных больших данных. Указываются те недостатки, которые приводят к снижению эффективности подразделения и/или организации в целом.

В пункте 3 дается обоснование необходимости применения технологий распределенной обработки данных для переработки выделенных больших данных. В этом пункте, с учетом специфики организации и выделенных в пункте 2 недостатков излагаются причины, вследствие которых целесообразно применение технологий распределенной обработки данных. Указываются те ожидаемые положительные результаты, которые будут получены в процессе применения технологий распределенной обработки данных.

В целом из материала отчета должно быть видно, что представляет собой организация и её основные процессы, с какой целью функционирует организация, какие задачи при этом решаются, какие большие данные циркулируют в организации, в чем причины недостаточно эффективной их переработки, почему и на каком уровне необходимо применение технологий распределенной обработки данных.

Контрольные вопросы.

1. Каковы цели и назначение организации?
2. Какова динамика информационных процессов, протекающих в организации?
3. Какие большие данные циркулируют в организации?
4. Чем обусловлено решение применения в организации технологий распределенной обработки данных?
5. Насколько рациональным является решение о применении технологий распределенной обработки данных для решения указанных в отчете задач организации?