# ВЫБОР ПРАВИЛЬНОЙ БАЗЫ ДАННЫХ ДЛЯ БОЛЬШИХ ДАННЫХ

*Ион ДОГА*

*Технический Университет Молдовы, Департамент Программной Инженерии и Автоматики*

***Аннотация:*** *В данной статье раскрывается базовые принципы по которым стоит выбирать базу данных для больших данных. Технология 3V является сегодня одной из самых актуальных. Приведены примеры баз данных.*

***Ключевые слова:*** *база данных, большие данные, конфиденциальность, big data.*

## 1. Введение в понятие BigData

В текущее время объемы информации растут экспоненциально. Для того чтобы быстрее реагировать на изменения рынка, получить конкурентные преимущества, повысить эффективность производства нужно получить, обработать и проанализировать огромное количество данных. Для работы с такими объемами информации инженеры были вынуждены модернизировать инструменты для работы над анализом всех данных. Так в 2000-х годах сформировалось понятие BigData, которое было интересно лишь узкому кругу специалистов. Сейчас это слово на слуху у любого, кто интересуется сферой информационных технологий. И это определение, а точнее направление развития ИТ, становится крайне популярным и стратегически важным в последнее время.

Технологии BigData позволяют обработать большой объем неструктурированных данных, систематизировать их, проанализировать и выявить закономерности там, где человеческий мозг никогда бы их не заметил. Это открывает совершенно новые возможности по использованию данных.

Само понятие BigData означает не просто большие пласты данных. Это огромные хранимые и обрабатываемые массивы из сотен гигабайт, и даже петабайт данных. Данных, которые можно обработать и извлечь из них некоторое количество полезной информации. Говоря коротко, можно определить BigData как совокупность технологий обработки информации для получения информации

## 2. Основные принципы и сферы применения BigData

Большие данные (big data) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence. В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях — весь мировой объём данных, и вытекающих из этого трансформационных последствий. В качестве определяющих характеристик для больших данных традиционно выделяют «три V»: объём (англ. volume, в смысле величины физического объёма), скорость (velocity в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов), многообразие (variety, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных. В дальнейшем возникли различные вариации и интерпретации этого признака. С точки зрения информационных технологий в совокупность подходов и инструментов изначально включались средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, системами управления базами данных категории NoSQL, алгоритмами MapReduce и реализующими их программными каркасами, и библиотеками проекта Hadoop. В дальнейшем, к серии технологий больших данных стали относить разнообразные информационно-технологические решения, в той или иной степени обеспечивающие сходные по характеристикам возможности по обработке сверхбольших массивов данных.

Важно заметить, объемы обрабатываемых через BigData данных постоянно растут, также, как и растет скорость ее обработки. Развитие этого направления вполне соответствует современному миру, стремительному и инновационному. С развитием BigData развивались и технологии, и наоборот. На текущий момент, BigData удел не только гигантов IT мира. Это направление, благодаря таким решениям как Hadoop от Apache Software Foundation, набору облачных сервисов от IBM, Amazon, Google, становится доступным практически любым компаниям, работающим в сфере ИТ. А такие

решения как Clickhouse, Cassandra, InfluxDB позволяют войти в сферу работы с BigData даже отдельным персонам.

Использование BigData на сегодняшний день становится обязательным условием для развития крупных ИТ компаний. Без анализа поведения своих пользователей, без возможности прогнозирования, руководствуясь только опытом и интуицией, уже крайне сложно оставаться конкурентоспособным. Настроенная и работающая система BigData способна в секунды предоставить ценнейшую информацию, полученную из анализа миллиардов действий клиентов компании. В текущем бизнесе уже зародилось понятие Data Driven Managment, которое подразумевает управление компанией исходя строго из анализа данных. И такие способы управления показывают блестящие результаты. Facebook, Google, Mail.ru, Яндекс уже давно используют аналитику для принятия решений. На сегодняшний момент в BigData заинтересован и традиционный бизнес, представители которого нуждаются в новых инструментах повышения эффективности.

Ниже представлены основные принципы работы с BigData.

- Горизонтальная масштабируемость: так как данных может быть много, то и система, в которой они хранятся должна быть расширяемой. Если объем данных вырос в 2 раза, то и количество кластеров увеличивается в 2 раза.
- Отказоустойчивость: горизонтальная масштабируемость подразумевает тот факт, что машин в кластере большое количество. И естественно эти машины будут по тем или иным причинам выходить из строя. К примеру, Hadoop-кластер Yahoo насчитывает более 42000 машин. Методы работы с BigData должны учитывать этот фактор и продолжать работать без видимых потерь.
- Локальность данных: в больших системах данные распределены на большом количестве машин. Если данные находятся на одной машине, а обрабатываются на другой, то расходы на передачу этих данных могут и вовсе превысить расходы на обработку. Поэтому важным вопросом в проектировании BigData стоит принцип локальности данных, обработке информации там же, где она хранится.

Сфера использования технологий BigData обширна. Так, с помощью BigData можно узнать о предпочтениях клиентов, об эффективности маркетинговых кампаний или провести анализ рисков. Ниже представлены результаты опроса IBM Institute, о направлениях использования BigData в компаниях.

Большинство компаний используют BigData в сфере клиентского сервиса, второе по популярности направление — операционная эффективность, в сфере управления рисками BigData менее распространены на текущий момент. Следует также отметить, что BigData являются одной из самых быстрорастущих сфер информационных технологий, согласно статистике, общий объем получаемых и хранимых данных удваивается каждые 1–2 года. За период с 2012 по 2014 год количество данных, ежемесячно передаваемых мобильными сетями, выросло на 81 %. По оценкам Cisco, в 2014 году объем мобильного трафика составил 2,5 эксабайта (единица измерения количества информации, равная 10^18 стандартным байтам) в месяц, а уже в 2019 году он будет равен 24,3 эксабайтам. Таким образом, BigData — это уже устоявшаяся сфера технологий, даже несмотря на относительно молодой ее возраст, получившая распространение во многих сферах бизнеса и играющая немаловажную роль в развитии компаний.

## 3. Основные решения

Технологии BigData, используемые для сбора и обработки BigData, можно разделить на 3 группы: программное обеспечение; оборудование; сервисные услуги. К наиболее распространенным подходам обработки данных относятся:

- SQL — язык структурированных запросов, позволяющий работать с базами данных. С помощью SQL можно создавать и модифицировать данные, а управлением массива данных занимается соответствующая система управления базами данных.
- NoSQL — термин расшифровывается как Not Only SQL (не только SQL). Включает в себя ряд подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать при постоянно меняющейся структуре данных. Например, для сбора и хранения информации в социальных сетях.
- MapReduce — модель распределения вычислений. Используется для параллельных вычислений над очень большими наборами данных (петабайты* и более). В программном

интерфейсе не данные передаются на обработку программе, а программа — данным. Таким образом запрос представляет собой отдельную программу. Принцип работы заключается в последовательной обработке данных двумя методами Map и Reduce. Map выбирает предварительные данные, Reduce агрегирует их.

- Hadoop — используется для реализации поисковых и контекстных механизмов высоконагруженных сайтов — Facebook, eBay, Amazon и др. Отличительной особенностью является то, что система защищена от выхода из строя любого из узлов кластера, так как каждый блок имеет, как минимум, одну копию данных на другом узле.
- SAP HANA — высокопроизводительная NewSQL платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что SAP HANA упрощает системный ландшафт, уменьшая затраты на поддержку аналитических систем.

## 4. Проблемы Bigdata

Проблемы системы BigData можно свести к трем основным группам: объем, скорость обработки, неструктурированность. Это три V–Volume, Velocity и Variety. Хранение больших объемов информации требует специальных условий, и это вопрос пространства и возможностей. Скорость связана не только с возможным замедлением и «торможением», вызываемым старыми методами обработок, это еще и вопрос интерактивности: чем быстрее процесс, тем больше отдача, тем продуктивнее результат. Проблема неоднородности и неструктурированности возникает по причине разрозненности источников, форматов и качества. Чтобы объединить данные и эффективно их обрабатывать, требуется не только работа по приведению их в пригодный для работы вид, но и определенные аналитические инструменты (системы). Но это еще не все. Существует проблема предела «величины» данных. Ее трудно установить, а значит трудно предугадать, какие технологии и сколько финансовых вливаний потребуется для дальнейших разработок. Ресурсы не бесконечны, хранение всех возможных данных в какой-то момент становится нецелесообразным. И встает необходимость отказа от части данных.

Собственно, это и является главной причиной отсрочки внедрения в компании проектов BigData (если не брать во внимание еще один фактор — довольно высокую стоимость). Подбор данных для обработки и алгоритм анализа может стать не меньшей проблемой, так как отсутствует понимание, какие данные следует собирать и хранить, а какие можно игнорировать. Становится очевидной еще одна «болевая точка» отрасли — нехватка профессиональных специалистов, которым можно было бы доверить глубинный анализ, создание отчетов для решения бизнес-задач и как следствие извлечение прибыли (возврат инвестиций) из BigData. Еще одна проблема BigData носит этический характер. А именно: чем сбор данных (особенно без ведома пользователя) отличается от нарушения границ частной жизни? Так, информация, сохраняемая в поисковых системах Google и Яндекс, позволяет им постоянно дорабатывать свои сервисы, делать их удобными для пользователей и создавать новые интерактивные программы. Поисковики записывают каждый клик пользователя в Интернете, им известен его IP-адрес, геолокация, интересы, онлайн-покупки, личные данные, почтовые сообщения и прочее, что, к примеру, позволяет демонстрировать контекстную рекламу в соответствии с поведением пользователя в Интернете. При этом согласия на это не спрашивается, а возможности выбора, какие сведения о себе предоставлять, не дается. То есть, по умолчанию в BigData собирается все, что затем будет храниться на серверах данных сайтов. Здесь можно затронуть другую проблему — обеспечение безопасности хранения и использования данных. Например, сведения о возможных покупателях и их история переходов на сайтах интернет-магазинов однозначно применимы для решения многих бизнес-задач. Но безопасна ли аналитическая платформа, которой потребители в автоматическом режиме (просто потому, что зашли на сайт) передают свои данные, — это вызывает множество споров. Современную вирусную активность и хакерские атаки не сдерживают даже супер-защищенные серверы правительственных спецслужб.

## Заключение

BigData открывает перед нами новые горизонты в планировании производства, образовании, здравоохранении и других отраслях. Если их развитие будет продолжаться, то технологии BigData могут поднять информацию, как фактор производства, на совершенно новый качественный уровень.

Информация станет не только равноценна труду и капиталу, но и возможно станет наиважнейшим ресурсом современной экономики.

**Библиография**


1. Веретенников А. В. BigData: анализ больших данных сегодня // Молодой ученый. — 2017. — №32. — С. 9-12. — URL https://moluch.ru/archive/166/45354/
2. Работа с Big Data: основные области и возможности . – [Электронный ресурс]. – Режим доступа: https://www.marketing.spb.ru/lib-around/stat/Big_Data.htm
3. Большие данные. – [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5