# Что такое «большие данные» (Big Data)

Резниченко В.А.

Институт программных систем НАН Украины

# Материальные и информационные технологии

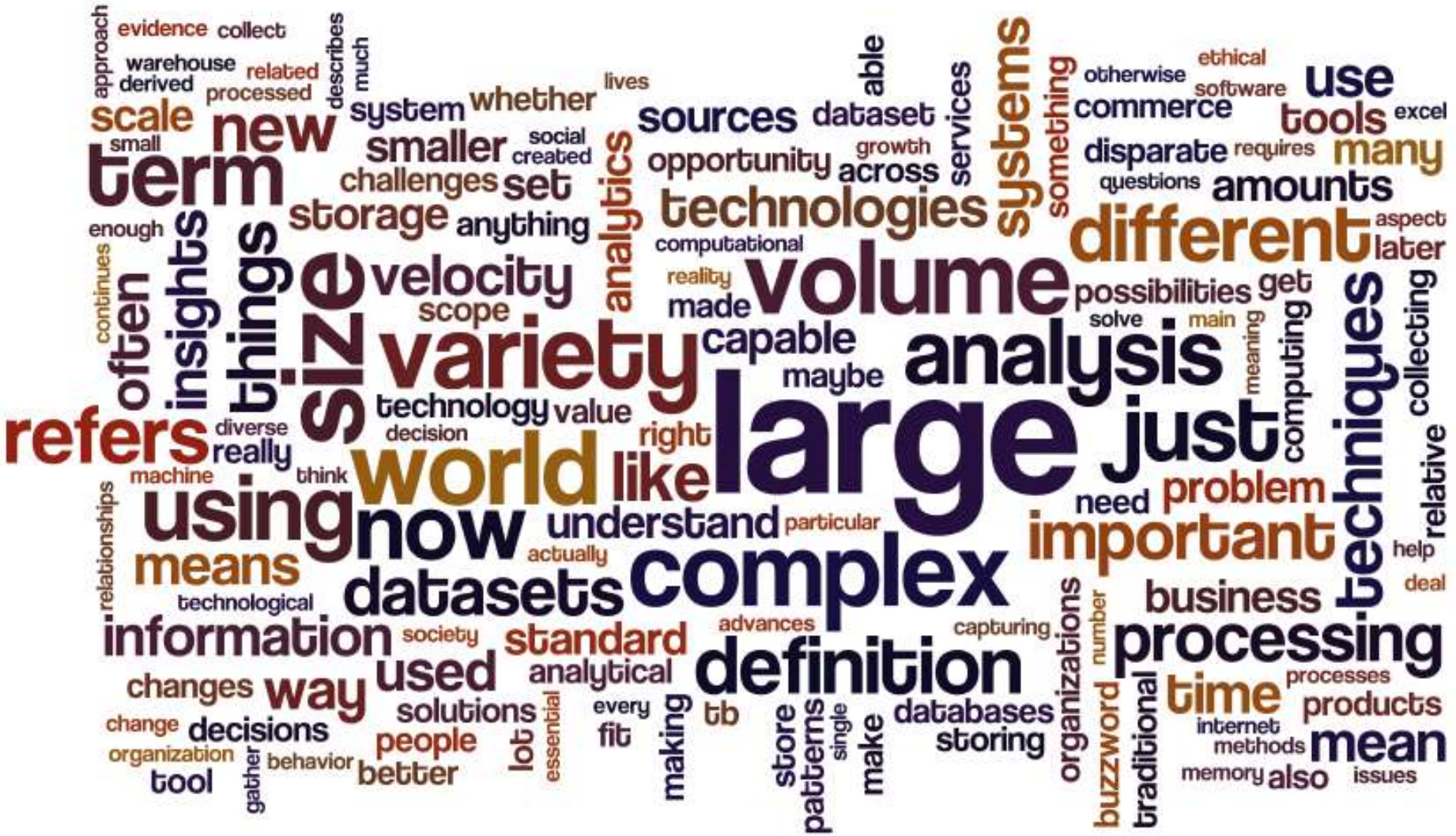
- К информационной технологии надо относиться как к материальной технологии. Практически все известные материальные технологии сводятся к процессу переработки, обработки или сборки специфического для них исходного сырья или каких-то иных компонентов с целью получения качественно новых продуктов.
- Логически информационные технологии мало чем отличаются от материальных технологий, на входе сырые данные, на выходе — структурированные, в форме, более удобной для восприятия человеком, данные, извлеченная из них информация, которая силой интеллекта (естественного или искусственного) превращается в полезное знание.
- На информационные технологии должны распространяться общие закономерности, согласно которых развиваются все остальные технологии, а это прежде всего **увеличение количества перерабатываемого сырья способствует повышению качества переработки**.

# Проблема больших объемов данных

- Проблема больших объемов данных вызвана не столько хлынувшим потоком данных, сколько нашей неспособностью старыми методами справиться с новыми объемами, вполне естественными для нынешнего этапа в развитии ИТ.

- Решение проблемы больших данных должно быть увязано с цепочкой «данные — информация — знание». Данные обрабатываются для получения информации, ее должно быть получено столько и в такой форме, чтобы человек или компьютер мог превратить информацию в знание, что, собственно и определяет методы работы с данными.

# Облако слов из определений BD



# Различные определения BD

- По адресу
- https://datascience.berkeley.edu/what-is-big-data/
- Дается 43 определения BD

# Определение больших данных (BD)

- **Big Data - BD** (большие данные) — огромные объемы неоднородной, неструктурированной или слабо структурированной, существенно распределенной и интенсивно растущей цифровой информации, которую невозможно обработать традиционными средствами, а также методы, технологии и средства их сбора, хранения, обработки и анализа с целью получения воспринимаемых человеком результатов.

# Единицы измерения данных и BD

| | | | |
|---|---|---|---|
| 1 байт | 8 бит | | |
| 1 килобайт (KB) | 1024 байт | $2^{10}$ байт | $10^{3}$ байт |
| 1 мегабайт (MB) | 1024 килобайт | $2^{20}$ байт | $10^{6}$ байт |
| 1 гигабайт (GB) | 1024 мегабайт | $2^{30}$ байт | $10^{9}$ байт |
| 1 терабайт (TB) | 1024 гигабайт | $2^{40}$ байт | $10^{12}$ байт |
| 1 петабайт (PB) | 1024 терабайт | $2^{50}$ байт | $10^{15}$ байт |
| 1 эксабайт (EB) | 1024 петабайт | $2^{60}$ байт | $10^{18}$ байт |
| 1 зетабайт (ZB) | 1024 эксабайт | $2^{70}$ байт | $10^{21}$ байт |
| 1 йоттабайт (YB) | 1024 зетабайт | $2^{80}$ байт | $10^{24}$ байт |
| 1 бронтобайт (BB) | 1024 йоттабайт | $2^{90}$ байт | $10^{27}$ байт |
| 1 геопбайт | 1024 бронтобайт | $2^{100}$ байт | $10^{30}$ байт |

# Определяющие характеристики BD-1

- Laney D (2001) 3d data management: Controlling data volume, velocity and variety. Technical Report 949, METAGroup (now Gartner). [Электронный ресурс]: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

- В 2001 г. Дуг Лани ведущий аналитик Gartner (ранее Meta Group) определили три признака данных большого объема (тогда еще НЕ было введено понятие BD), которые получили название **«Три V»: volume, velocity, variety (объем, скорость, многообразие)** и которые до сих пор считаются основополагающими признаками BD.

# Определяющие характеристики BD-2

- **Volume (объем)** — Мировой объем оцифрованной информации растет по экспоненте.
  - С 1980-х г. цифровая информация удваивается каждые 40 мес.
  - 2003 году мир накопил 5 эксабайтов данных а теперь это количество порождается каждые два дня.
  - К 2008 году этот объем вырос до 0,18 зеттабайта
  - 2011 — 1,76 зеттабайта,
  - 2013 — 4,4 зеттабайта.
  - 2015 — 6,5 зеттабайта.
  - 2020 (прогноз) — 40-44 зеттабайтов
  - 2025 — 163 зеттабайт.
  - В настоящее время площадь всех основных датацентров в мире равна площади 6000 футбольных полей.

Хранить информацию в одном месте бессмысленно и практически невозможно. Поэтому технология BD должна использовать распределенное хранение, управление, обработку и анализ данных, хранящихся в разнообразных хранилищах данных во всем мире.

# Определяющие характеристики BD-3

- **Velocity (скорость)** — Подразумевается как скорость прироста данных, так и скорость их обработки с целью получения конечных результатов.
  - Каждую минуту в мире посылается 204 миллиона е-писем,
  - Facebook: в день ставится 1.8 миллиона лайков, загружается 200 000 фото, загружается 100 ТВ данных.
  - Google обрабатывает в среднем 40 000 запросов в секунду (3,5 миллиарда в день), обрабатывает 25 петабайт в день.
  - В YouTube каждую минуту загружаются видео на 100 часов.
  - Если поместить в DVD все порожденные за день данные и положить эти диски друг на друга, то получится стопка, дважды превышающая расстояние до Луны.
  - По словам специалистов, к категории Big Data относится большинство потоков данных свыше 100 Гб в день.

# Определяющие характеристики BD-4

- **Variety (разноообразие, многообразие)** — Возможность воспринимать, хранить и обрабатывать различные данные, как с точки зрения их формата, так и структурирования. Традиционные базы данных позволяют хранить структурированные данные, но фактически в настоящее время порождаемые данные на 80% являются неструктурированными (тексты, видео, изображения, голос). Технология Big Data позволяет объединять и обрабатывать данные из различных источников и различных форматов.

# Определяющие характеристики BD-5

- Zikopoulos P, Parasuraman K, Deutsch T, Giles J, Corrigan D (2013) Harness the power of big data The IBM big data platform. McGraw Hill Professional, New York, NY. - [Электронный ресурс]: ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Harness_the_Power_of_Big_Data.pdf - расширение до **«5V»:**
  - **Value (ценность, значимость**). - Признак, описывающий экономический эффект, который технология обеспечивает пользователям. Например, по расчетам IBS, в 2013 году только 1,5% накопленных массивов данных имело информационную ценность.
  - **Veracity (достоверность**). Свойство, которое характеризует надежность данных. Является важной характеристикой Big Data в связи с тем, что исходные данные могут быть «сырыми» (неполными, нечеткими, расплывчатыми), и содержать много «шума». Технология Big Data учитывает этот фактор и позволяет надежно работать с такими данными.

# Определяющие характеристики BD-6

- Имеются предложения расширения этого списка до **«8V»:**

  - ***viability*** — жизнеспособность
  - ***variability*** — переменчивость
  - ***visualization*** — *визуализация*

- Наконец, позже к этим признакам добавили такое понятие, как **сложность (complexity)** поскольку управление данными может быть очень сложным, особенно при обработке больших объемов данных, которые приходят из разных источников.

# Классификация BD

- Dion Hinchcliffe. Big Data, The Moving Parts: Fast Data, Big Analytics, and Deep Insight. - https://www.flickr.com/photos/dionh/7550578346/in/photostream/

- Редактор журнала Web 2.0 Journal Дайон Хинчклифф (Dion Hinchcliffe) дал следующую классификацию BD:

  - быстрые данные (Fast Data) — тера-петабайты,
  - большая аналитика (Big Analytics) — пета-эксабайты
  - глубокое проникновение (Deep Insight) — экза-зеттабайты.

# Классификация BD- Fast Data

- Обработка данных типа Fast Data позволяет быстро воспринимать данные больших объемов и распознавать в них то, что вам нужно за приемлемое время. Она НЕ предполагает получения новых знаний, ее результаты соотносятся с априорными знаниями и позволяют судить о том, как протекают те или иные процессы. Это дает возможность лучше и детальнее увидеть происходящее, подтвердить или отвергнуть какие-то гипотезы.

- Только небольшая часть из существующих сейчас технологий подходит для решения задач Fast Data, в этот список попадают некоторые технологии работы с хранилищами данных, известные продукты Teradata, Netezza, Greenplum, СУБД типа Verica и kdb. Скорость работы этих технологий должна возрастать синхронно с ростом объемов данных.

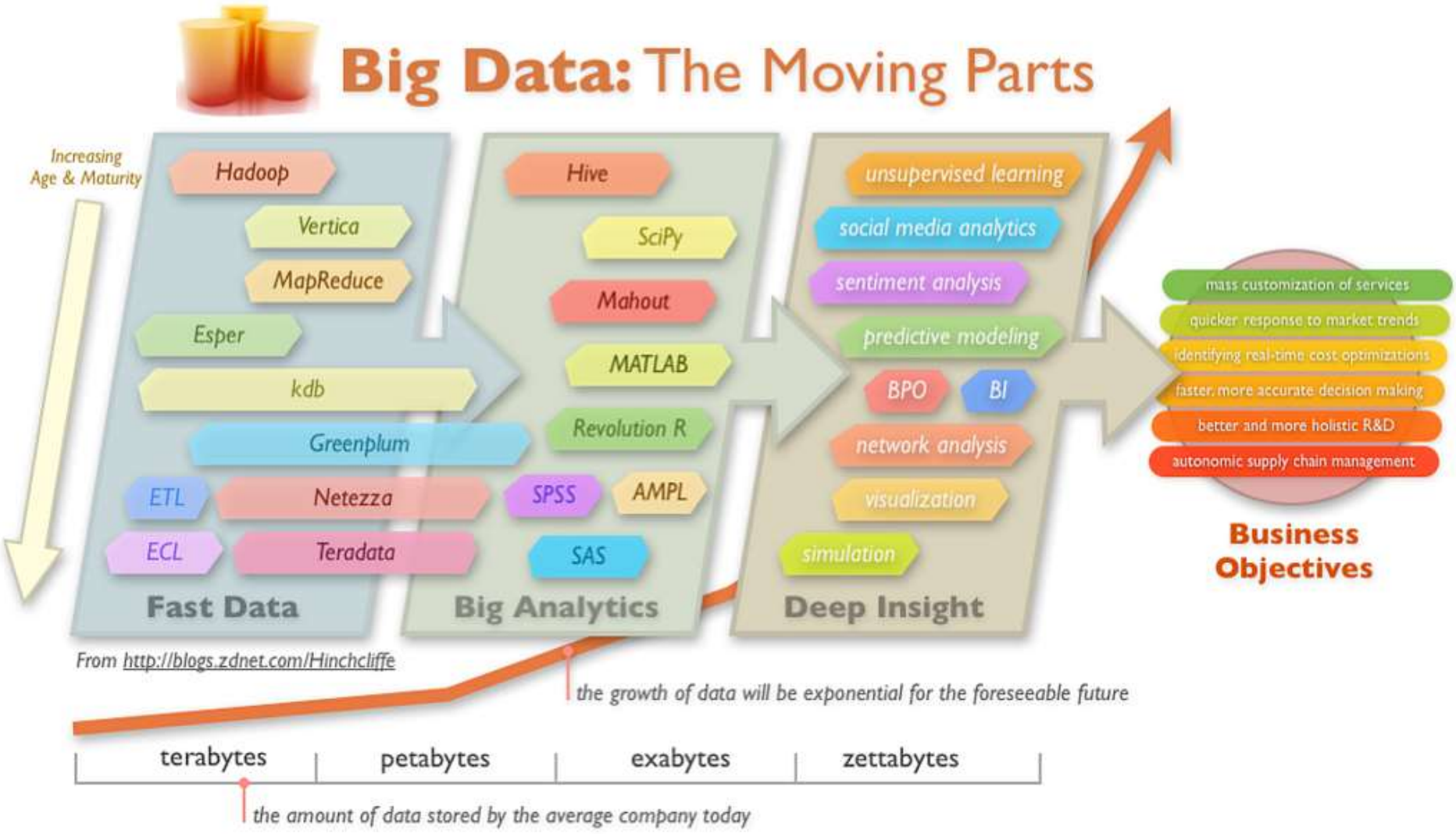
# Классификация BD- Big Analytics

- Задачи, решаемые средствами Big Analytics, отличаются от Fast Data, причем не только в количественном отношении, но и в качественном — соответствующие технологии должны помогать в получении новых знаний, служить для преобразования зафиксированной в данных информации в новое знание. Однако на этом уровне не предполагается наличие искусственного интеллекта при выборе решений или каких-либо автономных действий аналитической системы — она строится по принципу «обучения с учителем». Иначе говоря, весь ее аналитический потенциал должен быть заложен в нее в процессе обучения.
- Классическими примерами средств Big Analytics являются продукты MATLAB, SAS, Revolution R, более новые Apache Hive, SciPy Apache и Mahout, Netezza, Greenplum, СУБД типа Verica и kdb. Скорость работы этих технологий должна возрастать синхронно с ростом объемов данных.

# Классификация  BD- Deep Insight

- Высший уровень Deep Insight предполагает обучение без учителя (unsupervised learning) и использование современных методов аналитики (social media, predictive, segment, network и другие), а также различные способы визуализации. На этом уровне возможно обнаружение знаний и закономерностей, априорно неизвестных.

# Классификация BD



# История BD

- Клиффорд Линч (*Clifford Lynch*), редактор журнала Nature, подготовил в 2008 г. спецвыпуск «***Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных***?», в котором были собраны материалы о феномене взрывного роста объёмов и многообразия данных и технологических перспективах их использования.
- Введен в научной среде (проблема роста и многообразия научных данных, но начиная с 2009 года термин распространяется в бизнесе
- 2010 появляются первые продукты и решения, относящиеся обработке BD
- 2011 Крупнейшие ИТ-поставщики в своих деловых стратегиях начинают использовать понятие BD (IBM, Oracle, Microsoft, Hewlett-Packard, EMC)
- 2011 McKinsey «Большие данные: следующий рубеж в инновациях, конкуренции и производительности», оценила потенциальный рынок BD в миллиарды дол.
- 2011 компания Gartner отметила BD как тренд номер два в информационно-технологической инфраструктуре (после виртуализации)
- 2012 администрация президента США выделила 200 миллионов долларов для организации мероприятий по внедрению технологий BD в жизнь
- С 2013 года большие данные как академический предмет начинают изучать в появившихся вузовских программах по **науке о данных**.
- В 2015 году Gartner исключил BD из цикла зрелости новых технологий и прекратил выпускать выходивший в 2011—2014 годы отдельный цикл зрелости технологий BD .

# Методы и техники анализа, применимые к BD

- J. Manyika et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. - https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf
- **смешение и интеграция данных** (data fusion and integration) — методы, позволяющие интегрировать разнородные данные из разнообразных источников для возможности последующего глубинного анализа;
- методы класса **Data Mining**: обучение ассоциативным правилам (association rule learning), классификация, кластерный анализ, регрессионный анализ;
- машинное обучение;
- искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы;
- распознавание образов;
- прогнозная аналитика;
- имитационное моделирование;
- пространственный анализ (Spatial analysis) — класс методов, использующих топологическую, геометрическую и географическую информацию в данных;
- статистический анализ;
- визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

# Принципы работы с BD

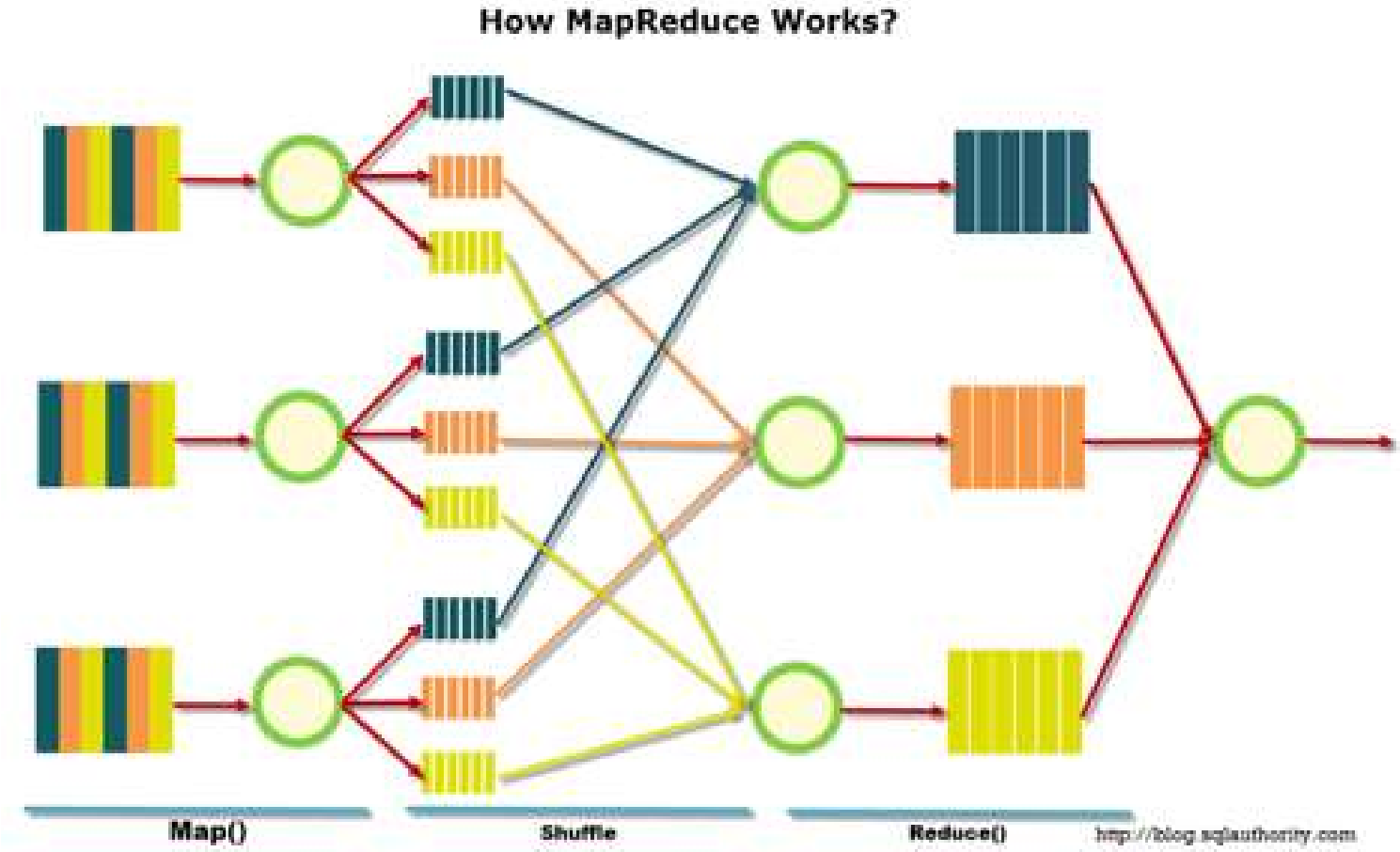
- **Горизонтальная масштабируемость**. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжает работать как и прежде.
- **Отказоустойчивость**. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин. Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий;
- **Локальность данных.** В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.
- Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы создания средств разработки данных. Одним из самых классических методов является MapReduce.

# Принципы работы с BD

- **Горизонтальная масштабируемость**. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжает работать как и прежде.
- **Отказоустойчивость**. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин. Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий;
- **Локальность данных.** В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.
- Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы создания средств разработки данных. Одним из самых классических методов является  MapReduce.

# MapReduce

- **MapReduce** – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах (большого количества  компьютерных узлов)




- MapReduce предполагает, что данные организованы в виде некоторых записей. Обработка данных происходит в 3 этапа: **Map, Shuffle, Reduce**

# Этапы MapReduce

- Этап **Map**. На этой стадии данные предобрабатываются при помощи функции map(), которую определяет пользователь. Работа этой стадии заключается в предобработке и фильтрации данных. Работа очень похожа на операцию map в функциональных языках программирования – пользовательская функция применяется к каждой входной записи.

  **Функция map() примененная к одной входной записи и выдаёт множество пар ключ-значение**. Что будет находиться в ключе и в значении – решать пользователю, но ключ – очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции reduce.

- Этап **Shuffle**. Проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. Эти корзины послужат входом для reduce.

- **Этап Reduce**. Каждая «корзина» со значениями, сформированными на этапе shuffle, попадает на вход функции reduce().

- **Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины».** Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

# Дополнительные факты про MapReduce

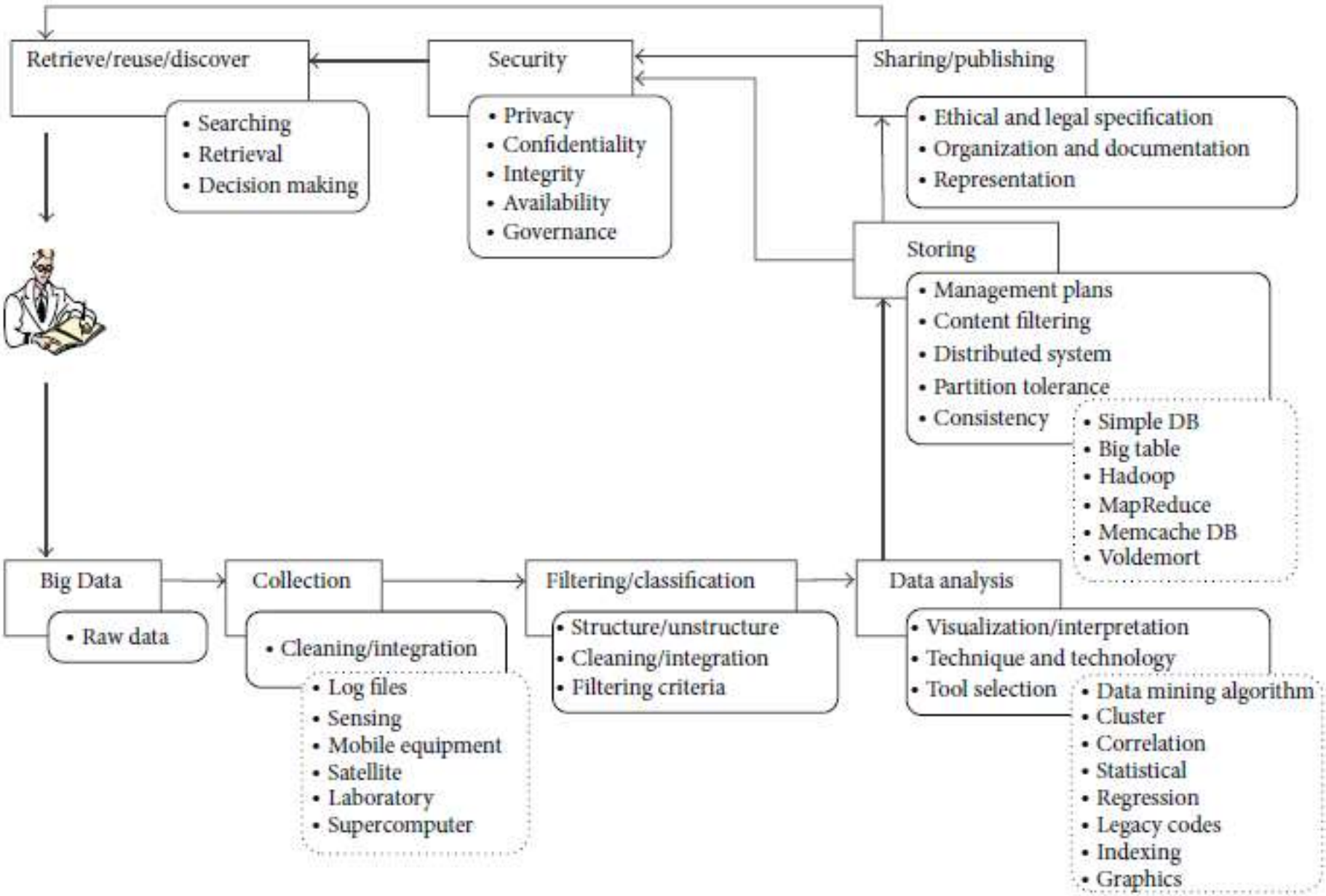
- Все запуски функции **map** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.

- Все запуски функции **reduce** работают независимо и могут работать параллельно, в том числе на разных машинах кластера.

- **Shuffle** внутри себя представляет параллельную сортировку, поэтому также может работать на разных машинах кластера. **Пункты 1-3 позволяют выполнить принцип горизонтальной масштабируемости.**

- Функция **map**, как правило, применяется на той же машине, на которой хранятся данные – это позволяет снизить передачу данных по сети (принцип локальности данных).

# Жизненный цикл BD



# Родственные технологии

- **Data lake** (озеро данных) — централизованное хранилище больших данных в сыром, необработанном виде. «Озера» хранят данные из разных источников, разных форматов, структурированные и неструктурированные. Они хранятся такими, как есть, без какой либо предварительной обработки.

- **Data science** (наука о данных) — дисциплина, изучающая проблемы анализа, обработки и представления информации в цифровой форме. 1974 год, Петер Наур «A Basic Principle of Data Science». В понятие data science входят все методы обработки оцифрованной информации и проектирования баз данных. Считают, что Data science включает в себя BD.

- **Business intelligence (бизнес-аналитика)** - совокупность методологий, процессов, архитектур и технологий, которые преобразуют большие объемы «сырых» данных в осмысленную и полезную информацию, пригодную для бизнес-анализа и для поддержки принятия оптимальных тактических и стратегических решений. 1958 г. статья исследователя из IBM Ханса Питера Луна