Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра интеллектуальных информационных технологий

Тэаро Кирилл Алексеевич

Data Quality

РЕФЕРАТ

Москва, 2022

Оглавление

1) Введение 2
2) Проблемы с данными в Data Science 3
3) Критерии качества больших данных 4
   - Доступность 5
   - Использование 6
   - Надежность 7
   - Актуальность 8
   - Представление 9
4) История 10
5) Data quality для ML 13
6) Способ оценки качества 14
   - Разбор оценки на примере 16
7) Методы повышения качества 21
8) Примеры из собственной научной работы 26
   - Chandra catalog 27
   - Pan-STARRS catalog 29
9) Вывод 32

Список источников 33

Введение

За последнее время изменилось множество технологий для отрасли IT.

XXI век стал настоящим прорывом для этой отрасли. Распространение глобального интернета, создание мобильный персональных компьютеров, развитие точности и удешевление разного рода датчиков – все это привело к экспоненциальному росту объёма данных, которые появляются в нашем мире в каждую секунду. Согласно данным компании Cisco (3 источник) в 2015 году интернет достиг 1 зеттабайта, а к 2025 году прогнозируется уже 175 зеттабайт. А есть еще множество данных, которые не входят в интернет, содержатся в хранилищах данных или просто не хранятся, а лишь обрабатываются. Сегодняшние обычные компьютеры способны хранить в себе от терабайта данных, что 20 лет назад было немыслимо. Это факт, что количество компьютерной информации с каждым годом растет огромными темпами (как минимум на примере интернета). И всю эту информацию надо как-то обрабатывать.

Обработка данных сегодня используются практически во всех областях. Медицина, экономика, бизнес, наука – это лишь одни из самых популярных областей использования. Но в последнее время стала развиваться аналитика для более далеких областей, таких как политика, игровая индустрия и даже кинематограф. И для всех этих отраслей нужны данные и нужны способы их обработки.

Но данные должны быть качественные, ведь составляя прогноз по ним мы ожидаем получить какую-то важную зависимость. А при плохих данных зависимость может получиться неверной и бизнес, например, потеряет деньги. Для решения всех этих важных проблемы создан отдельный подход по оценки Data Quality. Разные способы оценки, способы повышения качества и в целом применение Data Quality подходов в реальном мире будут рассмотрены в этом реферате.

Проблемы с данными в Data science

Аналитика больших данных относительно новая ветвь науки в мире. Соответственно сами большие данные сталкиваются с несколькими проблемами. Для их решения характеристику данных делят на 4 основных раздела: размер, скорость, разнообразие и ценности. Все эти характеристики для определенных данных могут быть улучшены разными способами. Например, размер можно увеличить добавлением новых данных, разнообразие – увеличением считывающих датчиков. А вот ценность данных напрямую зависит от их качества. Чем выше качество данных – тем важнее они для бизнеса или другой сферы. Компании стараются сами повышать качество данных, но сталкиваются со следующими проблемами:

1) Объем данных бывает огромен, и практически очень сложно оценивать их качество.
   Сегодня размер данных в мире увеличивается в два раза каждые 3 года. И становиться все сложнее собирать очищать и подготавливать качественные данные за приемлемое время. Сегодня это одна из главных проблем в повышении качества данных.

2) Разнообразие источников данных приводит к разнообразию типов данных и сложным структурам данных и увеличивает сложность интеграции данных.
   Приходится проводить дополнительную предобработку и каждой характеристики данных, приведению к одному типу. Это создает огромные проблемы так как с каждым годом разнообразие и качество устройств, считывающих данные растет, что вынуждает компании покупать новые модели. И совсем не всегда эти модели создают однотипные данные. Компаниям приходится либо обновлять все устройства сразу (что практически неосуществимо из-за дороговизны) либо стараться бороться с этой проблемой путем дополнительной предобработки данных.

3) Данные меняются быстро, а актуальность данных очень мала, что требует более высоких требований к технологии обработки.

   Компаниям очень важно получать актуальную информацию, ведь получившийся прогноз может быть неверен через определенное время. Обработка и анализ, основанные на устаревших данных, приведут к бесполезным выводам, что в конечном итоге приведет к ошибкам в принятии решений компании.

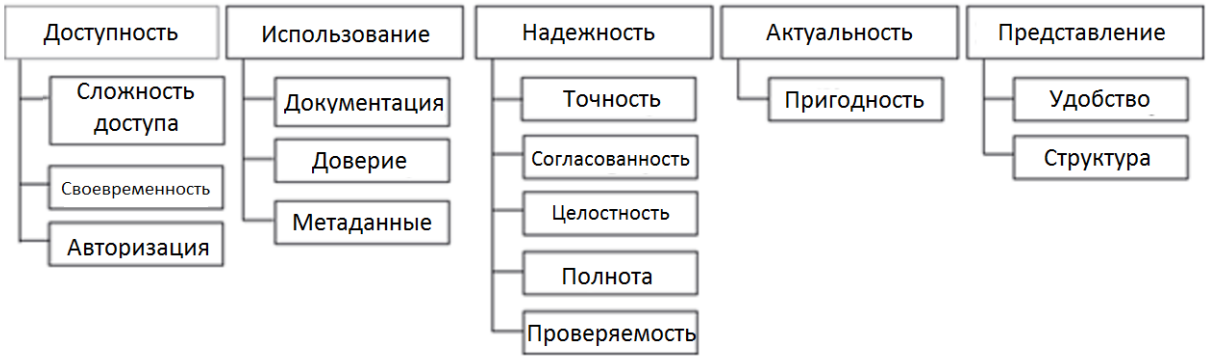
4) Сегодня в мире нет единых стандартов качества данных. Создание их только начинается.
   Чтобы гарантировать качество продукции и повысить выгоды для предприятий, в 1987 году Международная организация по стандартизации (ISO) опубликовала стандарты ИСО 9000 (8 источник). В настоящее время более 100 стран и регионов по всему миру активно внедряют эти стандарты. Это внедрение способствует взаимопониманию между предприятиями во внутренней и международной торговле и приносит пользу в виде устранения торговых барьеров. Напротив, изучение стандартов качества данных началось в 1990-х годах, но только в 2011 году ИСО опубликовала стандарты качества данных ISO 8000. В настоящее время более 20 стран приняли участие в этом стандарте, но по этому поводу существует много споров. Стандарты должны быть зрелыми и совершенными.

Критерии качества больших данных.

Большие данные — это новая концепция, и в мире еще не выработали единого определения качества данных и критериев качества. Одно можно сказать наверняка: качество данных зависит не только от их собственных характеристик, но и от бизнес-среды, использующей данные, включая бизнес-процессы и бизнес-пользователей. Только те данные, которые соответствуют соответствующим видам использования и отвечают требованиям, могут

считаться качественными данными. Обычно стандарты качества данных разрабатываются с точки зрения производителей данных. В прошлом потребителями данных были сами производители данных, что обеспечивало их качество. Однако в эпоху больших данных, при разнообразии источников данных, пользователи данных не обязательно являются производителями данных. Таким образом, очень трудно измерить их качество.

Можно представить общепринятые и широко распространённые показатели качества данных (1 источник), которые так или иначе оцениваются каждым бизнесом. Каждое из этих качеств делится на свои подкатегории исходя из специфики бизнеса. Таким образом, для оценки использовались иерархические стандарт качество больших данных.



Описание классического стандарта представлено далее.

1) Доступность
    - Сложность доступа
        - ♦ Предоставляется ли интерфейс доступа к данным?

        Доступность показывает уровень сложности получения данных. Она связана с открытостью данных, чем выше степень открытости данных, тем больше типов данных получается и тем выше степень доступности.

    - Своевременность
        - ♦ Поступают ли данные вовремя?
        - ♦ Регулярно ли обновляются данные?

- ♦ Соответствует ли требованиям временной интервал между получением данных до конца их обработки?

Своевременность определяется как задержка времени от генерации и сбора данных до их использования. Данные должны быть доступны в течение этой задержки, чтобы обеспечить возможность содержательного анализа. В век больших данных содержание данных постоянно меняется, поэтому своевременность имеет большое значение (11 источник).

- Авторизация

  Авторизация показывает, имеет ли пользователь право на использование данных.

2) Удобство использования

- Достоверность
  - ♦ Можно ли доверять организации/стране – источнику данных?
  - ♦ Как часто экспертам проводить аудит для проверки точности данных?
  - ♦ Приемлемые ли диапазоны значений у данных?

  Достоверность используется для оценки нечисловых данных. Это относится к объективным и субъективным компонентам достоверности источника или сообщения. Надежность данных зависит от трех ключевых факторов: надежности источников данных, нормализации данных и времени сбора данных.

- Документация

  Документация состоит из спецификации данных, которая включает в себя имя данных, определение, диапазоны допустимых значений, стандартные форматы, бизнес-правила и т.д. Нормативное определение данных повышает степень использования данных.

- Метаданные

  Поскольку потребители данных искажают значение общепринятой терминологии и концепций данных, с увеличением числа источников и типов данных использование данных может привести к рискам. Поэтому производителям данных необходимо предоставлять метаданные, описывающие различные аспекты

наборов данных, чтобы уменьшить проблемы, вызванные недопониманием или несоответствиями.

3) Надежность

- Точность
  - ♦ Насколько точные данные?
  - ♦ Хорошо ли представление данных отражает истинное состояние исходной информации?
  - ♦ Не вызовет ли предоставленная информация двусмысленность?

  Чтобы убедиться в точности заданного значения данных, оно сравнивается с известным эталонным значением. В некоторых ситуациях точность может быть легко измерена, например, пол, который имеет только два определенных значения: мужской и женский. Но в других случаях нет известного эталонного значения, что затрудняет измерение точности.

- Согласованность
  - ♦ Совпадают ли форматы и области значений после обработки данных, как и до обработки?
  - ♦ В течении определенного времени данные остаются непротиворечивыми и поддающимися проверке.

  Согласованность данных относится к тому, является ли логическая взаимосвязь между коррелированными данными правильной и полной. В области баз данных это обычно означает, что одни и те же данные, расположенные в разных областях хранения, должны рассматриваться как эквивалентные. Эквивалентность означает, что данные имеют равную ценность и одинаковое значение или, по существу, одинаковы. Синхронизация данных — это процесс обеспечения равенства данных.

- Целостность
  - ♦ Формат данных понятен и соответствует критериям?
  - ♦ Данные соответствуют структурной целостности?
  - ♦ Данные соответствуют целостности содержимого?

  Термин "целостность данных" имеет широкий охват и может иметь самые разные значения в зависимости от конкретного контекста. В

базе данных считается (13 источник), что данные с “целостностью” имеют полную структуру. Значения данных стандартизированы в соответствии с моделью данных и/или типом данных. Все характеристики данных должны быть правильными – включая бизнес-правила, взаимосвязи, даты, определения и т.д. В области информационной безопасности целостность данных означает поддержание и обеспечение точности и непротиворечивости данных на протяжении всего их жизненного цикла. Это означает, что данные не могут быть изменены несанкционированным или необнаруженным образом.

- Полнота
  - ♦ Повлияет ли недостаток компонента на использование данных для данных с несколькими компонентами?
  - ♦ Повлияет ли недостаток компонента на точность и целостность данных?

  Если исходное значение состоит из нескольких компонентов, мы можем описать качество с полнотой. Полнота означает, что значения всех компонентов одного элемента данных являются действительными. Например, для цвета изображения RGB может использоваться для описания красного, зеленого и синего цветов, а RGB представляет все части цветовых данных. Если значение цвета определенного компонента отсутствует, изображение не может отображать реальный цвет, и его полнота нарушается.
- Проверяемость

  С точки зрения применения аудита жизненный цикл данных включает в себя три фазы: генерацию данных, сбор данных и использование данных. Но здесь проверяемость означает, что аудиторы могут справедливо оценить точность и целостность данных в рамках разумных временных и трудовых ограничений на этапе использования данных.

4) Актуальность
   - Пригодность

- Собранные данные не полностью соответствуют теме, но они раскрывают один аспект
- Большинство извлеченных наборов данных относятся к теме поиска, необходимой пользователям?
- Совпадает ли информационная тема с темой поиска пользователей

Пригодность имеет двухуровневые требования: 1) объем доступных данных, используемых пользователями, и 2) степень, в которой полученные данные соответствуют потребностям пользователей в аспектах определения показателей, элементов, классификации и т. д.

5) Качество представления
   - Удобство чтения
     - Данные (содержание, формат и т.д.) ясны и понятны?
     - Понятно ли, что предоставленные данные соответствуют потребностям?
     - Описание данных, классификация и содержание кодирования соответствуют спецификации и легки для понимания?

     Удобочитаемость определяется как способность содержимого данных быть правильно объясненным в соответствии с известными или четко определенными терминами, атрибутами, единицами измерения, кодами, сокращениями или другой информацией.
   - Структура

     Более 80% всех данных неструктурированные, поэтому структура относится к уровню сложности преобразования полу структурированных или неструктурированных данных в структурированные данные с помощью технологии.

Конечно, для каждой структуры бизнеса характеристики качества будут отличаться, однако они в любом случае будут близки к представленной выше классификации.

История Data quality

Зарождение Data quality

Еще в 1865 году профессор Ричард Миллар Девенс ввел термин “бизнес-аналитика” (сокращенно BI) в своей циклопедии коммерческих и бизнес-анекдотов. Он использовал этот термин, чтобы описать, как сэр Генри Фернезе собирал информацию, а затем действовал на ее основе раньше, чем это сделали его конкуренты, чтобы увеличить свою прибыль.

Гораздо позже, в 1958 году, Ханс Петер Лун написал статью, описывающую потенциал сбора BI с помощью технологии. Современная версия Business Intelligence используют различные технологии сбора, анализа данных и преобразования их в полезную информацию. Затем эта информация используется, чтобы обеспечить значительное преимущество для бизнеса. По сути, современная бизнес-информация ориентирована на использование технологий для быстрого и эффективного принятия обоснованных решений.

В 1968 году люди со специализированными навыками были единственными, кто мог преобразовать имеющиеся данные в полезную информацию. В то время данные, взятые из нескольких источников, обычно хранились изолированно. Исследование такого рода данных обычно включало работу с фрагментированной, разрозненной информацией и приводило к получению сомнительных отчетов. Эдгар Кодд осознал эту проблему и представил решение в 1970 году, которое изменило представление людей о базах данных. Его решение предполагало создание “модели реляционной базы данных”, которая приобрела огромную популярность и была принята во всем мире.

Система управления базами данных

Системы поддержки принятия решений (DSS) описываются как самая ранняя система управления базами данных. Многие историки предполагают, что современная бизнес-аналитика основана на базе данных DSS. В 1980-х годах число поставщиков BI существенно выросло. Деловые люди открыли для себя ценность больших данных и современной бизнес-аналитики. За это время был создан и развит широкий спектр инструментов, ориентированных на цели доступа к данным и их организации более эффективными и простыми способами. Исполнительные информационные системы и хранилища данных являются

примерами некоторых разработанных инструментов. Важность качества данных помогла стимулировать разработку реляционных баз данных.

Data quality как сервис

В 1986 году, до появления недорогого хранилища данных, поддерживались огромные мэйнфреймовые компьютеры, которые содержали данные об имени и адресе, используемые для служб доставки. Это позволило перенаправить почту по назначению. Эти мэйнфреймы были разработаны для исправления распространенных орфографических ошибок в именах и адресах, а также для отслеживания клиентов, которые умерли, переехали, попали в тюрьму, развелись или вступили в брак. Это было также время, когда правительственные учреждения предоставили почтовые данные “сервисным компаниям” для перекрестных ссылок с реестром NCOA (National Change of Address). Это решение сэкономило нескольким крупным компаниям миллионы долларов, поскольку больше не было необходимости в ручном исправлении данных клиентов и удалось избежать ненужных почтовых расходов. Эта ранняя попытка повысить точность /качество данных изначально продавалась как услуга.

Неиссякаемый поток данных

В конце 1980-х и начале 1990-х годов многие организации начали осознавать ценность данных и интеллектуального анализа данных. Руководители компаний и лица, принимающие решения, все больше полагались на анализ данных. Кроме того, бизнес-процессы создавали все большие и большие объемы данных из разных отделов для разных целей. Затем, вдобавок ко всему, интернет стал популярным. В 1990-х годах Интернет стал чрезвычайно популярным, и реляционные базы данных, принадлежащие крупным корпорациям, не могли поспевать за огромным потоком доступных им данных. Эти проблемы усугублялись разнообразием типов данных и нереляционных данных, которые развивались в течение этого времени. Нереляционные базы данных, часто называемые NoSQL, появились как решение. Базы данных NoSQL могут быстро переводить различные типы данных и позволяют избежать жесткости баз данных SQL, устраняя “организованное” хранилище и предлагая большую гибкость.

Нереляционные базы данных были разработаны в ответ на данные из Интернета, необходимость обработки неструктурированных данных и стремление к более быстрой обработке. Модели NoSQL основаны на распределенной системе

баз данных, использующей несколько компьютеров. Нереляционные системы работают быстрее, организуют данные с использованием специального подхода и обрабатывают значительные объемы различных типов данных. Для общих исследований NoSQL является лучшим выбором при работе с большими неструктурированными наборами данных (big data), чем реляционные базы данных из-за их скорости и гибкости. Термин “большие данные” стал официальным в 2005 году.

Управление данными

К 2010 году объем и сложность данных продолжали расти, и в ответ предприятия стали более изощренными в использовании данных. Они разработали методы объединения, манипулирования, хранения и представления информации. Это было началом управления данными.

Дальновидные компании создали управляющие организации для хранения бизнес-данных и разработали совместные процессы для использования данных, необходимых для бизнеса. Но что более важно, они разработали “политико-ориентированный подход” к стандартам качества данных, моделям данных и безопасности данных. Эти ранние группы игнорировали представления о все более крупных и сложных хранилищах и сосредоточились на политиках, которые определяли, внедряли и обеспечивали соблюдение интеллектуальных процедур для данных. Одна процедура позволяет хранить один и тот же тип данных в нескольких местах при условии соблюдения одних и тех же политик. В результате предприятия брали на себя все больше и больше ответственности за содержание своих данных. В настоящее время данные широко признаны ценным корпоративным активом.

Программа эффективного управления данными организовала руководящий орган из хорошо информированных людей и разработала меры реагирования на различные ситуации. Поведение по управлению данными должно быть четко определено, чтобы эффективно объяснить, как данные будут обрабатываться, храниться, создаваться резервные копии и в целом защищаться от ошибок, кражи и атак. Должны быть разработаны процедуры, определяющие, как должны использоваться данные и каким персоналом. Кроме того, необходимо внедрить набор средств контроля и аудиторских процедур, которые обеспечивают постоянное соблюдение внутренней политики в отношении данных и внешних

правительственных постановлений, а также гарантируют согласованное использование данных во множестве корпоративных приложений. Машинное обучение стало популярным способом реализации управления данными.

Data quality для задач машинного обучения.

На сегодняшний день качество многих данных для машинного обучение невысоко, что приводит к аналитическим ошибкам. В своей статье Harvard Business Review (HBR) “Если ваши данные плохие, ваши инструменты машинного обучения бесполезны” (7 источник) Томас К. Редман так резюмирует текущую проблему качества данных: “Все более сложные проблемы требуют не просто большего количества данных, но и более разнообразных, всеобъемлющих данных. А вместе с этим возникает еще больше проблем с качеством”. В другой статье HBR Редман отмечает, что, по оценкам IBM, некачественные данные обходятся предприятиям в 3,1 триллиона долларов в год только в США (15 источник). Для любой компании, которая хочет участвовать в революции машинного обучения, которая уже разрушает многие аспекты современного бизнес-ландшафта, качество данных — это проблема, которой просто невозможно избежать.

Способ оценки качества

Определение целей сбора данных является первым шагом всего процесса оценки. Пользователи больших данных рационально выбирают данные для использования в соответствии со своими стратегическими целями или бизнес-требованиями, такими как операции, принятие решений и планирование. Источники данных, типы, объем, требования к качеству, критерии оценки и спецификации, а также ожидаемые цели должны быть определены заранее.

В разных бизнес-средах выбор элементов качества данных будет отличаться. Например, для данных социальных сетей своевременность и точность являются двумя важными качественными характеристиками. Однако, поскольку трудно непосредственно судить о точности для оценки исходных данных необходима некоторая дополнительная информация, а другие источники данных служат дополнениями или доказательствами. Таким образом, доверие стало важным аспектом качества. Однако данные социальных сетей обычно неструктурированные, и их последовательность и целостность не подходят для оценки. Область биологии является важным источником больших данных. Однако из-за отсутствия единых стандартов программное обеспечение для хранения данных и форматы данных сильно различаются. Таким образом, трудно рассматривать согласованность как аспект качества, а потребности в отношении своевременности и полноты как аспектов качества данных невелики.

Для дальнейшей оценки качества необходимо выбрать конкретные показатели оценки для каждого измерения. Они требуют, чтобы данные соответствовали определенным условиям или функциям. Формулировка оценочных показателей также зависит от фактической бизнес-среды.

Для каждого измерения качества требуются различные инструменты, методы и процессы измерения, что приводит к различиям во времени оценки, затратах и человеческих ресурсах. При четком понимании работы, необходимой для оценки каждого измерения, выбор тех измерений, которые отвечают потребностям, может хорошо определить масштаб проекта. Предварительные результаты оценки параметров качества данных определяют базовую линию, в

то время как остальная оценка как часть бизнес-процесса используется для непрерывного обнаружения и улучшения информации.

После завершения подготовки к оценке качества процесс переходит в фазу сбора данных. Существует множество способов сбора данных, в том числе: интеграция данных, поиск-загрузка, веб-сканеры, методы агентов, мониторы операторов и т.д. В эпоху больших данных сбор данных относительно прост, но большая часть собранных данных не всегда хороша. Нам необходимо улучшить качество данных, насколько это возможно в этих условиях, без значительного увеличения стоимости сбора.

Источники больших данных очень широки, а структуры данных сложны. Полученные данные могут иметь проблемы с качеством, такие как ошибки в данных, недостающая информация, несоответствия, шум и т.д. Целью очистки данных (data scrubbing) является обнаружение и удаление ошибок и несоответствий из данных с целью улучшения их качества. Очистка данных может быть разделена на четыре модели, основанные на методах реализации и областях применения: ручная реализация, написание специальных прикладных программ, очистка данных, не связанных с конкретными областями применения, и решение проблемы типа конкретной прикладной области. Из этих четырех подходов третий имеет хорошую практическую ценность и может быть успешно применен.

Затем процесс переходит к этапам оценки качества данных и мониторинга. Суть оценки качества данных заключается в том, как оценивать каждое измерение. Текущий метод делится на две категории: качественные и количественные методы. Метод качественной оценки основан на определенных критериях оценки и требованиях, в соответствии с целями оценки и запросами пользователей, с точки зрения качественного анализа для описания и оценки ресурсов данных. Качественный анализ должен проводиться предметными экспертами или профессионалами. Количественный метод — это формальный, объективный и систематический процесс, в котором для получения информации используются числовые данные. Таким образом, объективность, обобщаемость и цифры — это характеристики, часто ассоциирующиеся с этим методом, результаты оценки которого более интуитивны и конкретны.

После оценки данные можно сравнить с исходным уровнем для оценки качества данных, установленным выше. Если качество данных соответствует базовому стандарту, можно перейти к последующей фазе анализа данных, и будет сгенерирован отчет о качестве данных. В противном случае, если качество данных не соответствует базовому стандарту, необходимо получить новые данные.

Строго говоря, анализ данных и интеллектуальный анализ данных не относятся к сфере оценки качества больших данных, но они играют важную роль в динамической корректировке и обратной связи при оценке качества данных. Можно использовать эти два метода, чтобы выяснить, существует ли ценная информация или знания в больших данных и могут ли эти знания быть полезны для политических предложений, бизнес-решений, научных открытий, лечения заболеваний и т.д. Если результаты анализа соответствуют цели, то результаты выводятся и передаются обратно в систему оценки качества, чтобы обеспечить лучшую поддержку для следующего раунда оценки. Если результаты не достигают цели, базовый уровень оценки качества данных может оказаться необоснованным, и необходимо своевременно скорректировать его, чтобы получить результаты в соответствии с нашими целями.

Рассмотрим один из способов определения с низкой достоверностью данных на примере.

Определение качества на примере

Пусть есть устройства считывания данных, состоящее из 4 подпроцессов, которые в последовательном доступе обрабатывают информацию и наносят ее в базу данных.

Схема:

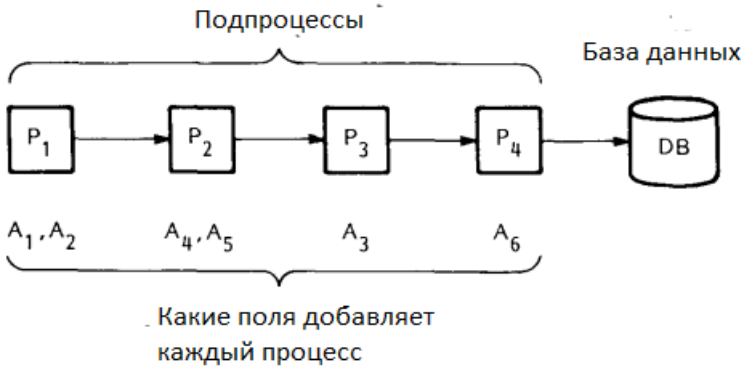


Рис 1

Процессы передают нам такую информацию об одном объекте:

	P_1	P_2	P_3	P_4
A_1	XYZ1	XYZ1	XYZ1	**XYZ-001**
A_2	Yes	Yes	**No**	No
A_3			K	K
A_4		1500	**5100**	5100
A_5		On	On	**1**
A_6				OK
Date	3/1/89	3/2/89	3/20/89	3/25/89

Рис 2

Способ оценки качества оборудования, а также достоверность данных будет основываться на случайной выборке объектов из всего множества. Рассмотрим на примере Рис 2.

Отслеживание данных обсуждается в контексте процесса сбора информации, хотя концептуально эти идеи аналогичным образом применимы и к другим процессам. Различные поля собираются из нескольких источников в течение определенного периода времени, и запись данных формируется путем объединения соответствующих частей данных, которые затем сохраняются в базе данных. Из-за его сходства с процессом сборки в производстве этот

процесс будет называться процессом сборки информации. Качество базы данных критически зависит от качества этого процесса сбора информации.

Ключевыми шагами в применении отслеживания данных являются следующие:

1) Выбор случайной выборки записей, поступающих в процесс.
2) Отслеживание отобранные записи по мере их прохождения через процесс. Чтобы сделать это, необходимо записать содержимое в каждой отобранной записи по мере выхода из каждого подпроцесса и время выхода. Рисунок 2 представляет собой пример записи, которая была введена в процесс 1 марта 1989 года и внесена в базу данных 25 марта 1989 года.
3) Нужно определить ошибки, вызванные процессом и составляющими его подпроцессами.
4) Через соответствующие промежутки времени суммировать прогресс выборочных записей в процессе. Нужно разрабатывать соответствующие графики и сводки для проверки соответствия стандартам и анализа изменений или ошибок

Данный способ обеспечивает:

- Постоянное измерение уровня качества процесса
- Постоянный контроль процесса
- Возможности для постоянного совершенствования процесса путем выявления подпроцессов, которые вносят ошибки или несоответствия

В следующих подразделах представлена более подробная информация.

Шаг 1. Выборка.

Выборка используется так, как она имеет много преимуществ по сравнению с полной проверкой. Многие процессы сбора информации являются "пакетными" в том смысле, что группы записей входят в процесс вместе. В этом случае уместна случайная выборка из каждой партии. Если записи поступают "по одной за раз", возможно, более целесообразно отбирать каждую запись независимо от других с заранее определенной вероятностью отбора.

Шаг 2. Отслеживание.

Отобранным записям присваивается уникальное идентификационное поле, чтобы их можно было легко отслеживать с момента их включения в процесс до внесения в базу данных. Это очень важно, поскольку определенные изменения в полях при переходе от одного подпроцесса к другому используются в качестве индикаторов ошибок в процессе. Рисунок 2 представляет собой пример конкретной записи данных, отобранной и отслеженной в ходе процесса. Измененные поля данных выделены жирным. Эти изменения могут быть классифицированы следующим образом:

- Нормализация: изменения, такие как вставка или удаление разделителей, пробелов и т.д., для соответствия различным критериям формата, используемым на протяжении всего процесса.
- Перевод: изменения из-за использования разных языков на протяжении всего процесса.
- Ложно-операционные: изменения, которые происходят, когда один подпроцесс изменяет элемент данных на другое значение. Ложные операционные изменения указывают на ошибку где-то в процессе.

На рисунке 2 показано изменение нормализации для поля данных A1 на этапе P4, а изменение перевода показано для поля данных A5 на этапе P4. Хотя изменения в нормализации и переводе не обязательно являются ошибками, предпочтительно минимизировать эти изменения везде, где это возможно. Это может быть сделано путем внедрения словарей данных, которые предоставляют стандартизированные форматы или языки в разных системах.

Ложно-операционные изменения являются наиболее серьезным типом изменений. В таблице 1 показаны изменения, связанные с ложными операциями, для поля данных A2 на этапе P3 и для поля данных A 4 на стадии P3. Анализ этих изменений, происходящих с течением времени, может привести к выявлению этапов процесса, которые вносят наиболее существенные ошибки или несоответствия в базу данных.

Шаг 3. Определение ошибок.

Как отмечалось выше, изменения в полях могут быть использованы в качестве индикаторов ошибок процесса. Большинство ошибок легко

выявляются, хотя некоторые могут потребовать дальнейшей работы. В таких случаях иногда полезно сравнить выборочные записи с соответствующим стандартом чтобы убедиться, что все ошибки были выявлены. Важно отличать аудит данных, уже имеющихся в базе данных, от отслеживания данных. Цели аудита данных, уже имеющихся в базе данных, заключаются в выявлении ошибок и оценке частоты ошибок, в то время как цели отслеживания данных заключаются в выявлении мест возникновения ошибок и предотвращении дальнейших ошибок.

В результате анализа 100 записей получилась такая статистика:

Рисунок 3

Поле	Этап	Изменений	Записей	Процент
A_2	P_2	3	100	3%
A_2	P_3	21	94	22%
A_2	P_4	2	91	2%
A_3	P_4	1	91	1%
A_4	P_3	4	94	4%
A_4	P_4	7	91	8%

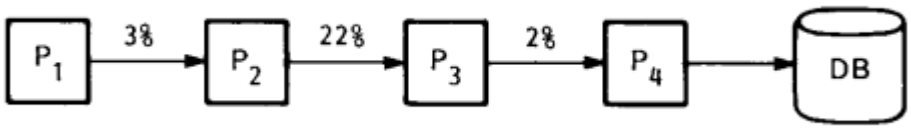


Рисунок 4

Шаг 4. Оценка качества.

Через соответствующие промежутки времени следует составлять резюме хода обработки выборочных записей в рамках процесса. Сводные статистические данные следует сопоставлять с целями и соответствующими действиями, предпринятыми для понимания и устранения коренных причин расхождений. Кроме того, для улучшения качества данных необходимо изменить сам процесс. Сводная статистика и более глубокий анализ отслеживания данных могут помочь предложить соответствующие модификации процесса.

Можно получить следующие результаты оценки качества данных:

1) Рисунок 3 полезен при выявлении качества достоверности полей данных. В текущем примере видно, что качество поля A2 плохое.
2) Рисунок 4 показывает для поля A2 на каких этапах происходят ошибки. В данном примере видно, что изменения происходят в основном на этапе P3. Следует искать первопричины таких изменений, чтобы добиться наиболее значительных улучшений в процессе.

В то время как отслеживание данных предоставляет данные, необходимые для измерения, контроля и повышения уровня качества процесса, фактический контроль и улучшение зависят от выявления причин ошибок и устранения этих причин. Для этого требуется активный "владелец процесса", несущий ответственность как за то, чтобы процесс работал хорошо, так и за полномочия изменять процесс по мере необходимости.

Таким образом, можно на практике определять и оценивать качество достоверности полученных данных.

Далее можно рассмотреть методы повышения качества данных. А также важный аспект увеличения выборки данных.

Методы повышения качества

Все методы повышения качества можно разделить на 3 группы:

1) Очистка данных - процесс выявления и исправления ошибок в исходной информации, т. е. оценка достоверности данных, выявление ошибочных подозрительных данных: аномалий, дубликатов, противоречий и т. п.
2) Предобработка данных – процесс подготовки данных к решению конкретной задачи и приведение их в соответствие с требованиями, определяемыми этой задачей и способами ее решения, т. е. понижение размерности исходной информации, устранение незначащих признаков и т. п.
3) Обогащение данных – процесс насыщения данных новой информацией, позволяющей сделать их более ценной для определенной аналитической

задачи, т. е. привлечение информации из других источников, заполнение пропусков в информации, выявление связей между объектами и т. п.

В то же время, если методы очистки и предварительной обработки данных можно целиком отнести к одному из этапов так называемого процесса ETL (extraction, transformation, loading - извлечение, преобразование, загрузка) (18 источник), то со способами обогащения данных такой однозначности нет. Действительно, например, идентификация связей между объектами связана с обработкой данных, уже загруженных в хранилище данных (CD), и предусматривает получение полезной информации, которая не доступна явно, но может быть получена путем манипулирования существующими данными. Затем эта информация внедряется в виде новых полей или таблиц и может быть использована в дальнейшем анализе.

В этой связи представляется обоснованным разделять обогащение данных на два вида – внешнее и внутреннее.

Внешнее обогащение данных связано с решением бизнес-задач, требующих повышенного качества работы. Именно в этом случае необходимо в распоряжение аналитиков организации привлекать дополнительную информацию из других источников с тем, чтобы обогатить внутренние данные до уровня информативности, который позволит с высоким качеством решать задачи. К внешним источникам данных можно отнести: другие организации, работающие в схожей сфере; органы государственной власти и местного самоуправления, включая налоговые и статистические службы; финансово-кредитные учреждения, банки, страховые компании; службы социальной сферы, включая органы труда и занятости, систему здравоохранения, пенсионный фонд.

Внутреннее обогащение данных не нуждается в привлечения внешней информации, поскольку повышение информативности данных достигается путем изменения их организации.

Важно понимать, что применение любого метода или комплекса методов повышения качества данных, к какой бы группе или виду они не относились, требует предварительной оценки качества данных с целью выявления характерных проблем и уровня их сложности, а также выработки

соответствующей стратегии по их решению. Поэтому на предприятиях появилась новая дисциплина – управление качеством данных на предприятии (Enterprise Data Quality Management, EDQM). Более того, EDQM стало частью общего процесса управления качеством на предприятии

Ключевым звеном EDQM является именно оценка качества данных, которая реализуется на основе единовременной оценки, мониторинга или визуальной оценки. В любом случае разработка методики оценки качества данных требует ответа на вопрос: где именно ее следует проводить? При этом следует рассматривать следующие варианты: непосредственно в источниках данных, в ETL-процессе и в аналитической системе.

1) Первый из этих вариантов, т. е. оценка качества данных непосредственно в источниках данных, позволяет эффективно выполнить поиск орфографических ошибок, пропущенных, аномальных, логически неверных и фиктивных значений, противоречий и дубликатов на уровне записей и таблиц. Преимущества данного варианта в том, что результаты оценки качества данных, определенные методы очистки данных могут быть задействованы уже в ETL-процессе и в ХД поступят очищенные данные. Но надо помнить, что в ходе ETL-процесса качество данных может вновь ухудшится, поскольку происходит интегрирование данных из нескольких источников и могут появиться новые дубликаты и противоречия, несоответствия форматов и т. д. Следовательно, записи, уникальные и непротиворечивые для одного источника, могут потерять уникальность и непротиворечивость после объединения или слияния источников.
2) Второй вариант, т. е. оценка качества данных в ETL-процессе, в соответствии с выявленными проблемами и итогами оценки качества данных позволяет оперативно задействовать методы их очистки, загружая в ХД уже достоверную информацию. Хотя при этом возникает другая проблема, обусловленная тем, что использование данного подхода может заметно увеличить так называемое загрузочное окно, в течение которого возрастает нагрузка на информационную систему организации.
3) В третьем случае, т. е. оценки качества данных в аналитической системе, а именно в процессе предобработки данных перед применением к ним различных методов Data Mining, эта оценка производится аналитиком

визуально с использованием таблиц и диаграмм, а также на основе статистических оценок и характеристик. Действительно, с помощью гистограмм, например, легко можно выявить аномальные значения, а оценка дисперсии позволяет оценить степень неравномерности ряда значений.

Конечно, наиболее эффективным решением является использование всех трех вариантов, однако факторы времени и трудозатрат далеко не всегда позволяют выбирать именно это. В любом случае не стоит забывать, что цель оценки качества данных – это лишь выявление в них каких-либо проблем, а локализация источников этих проблем и, тем более, борьба с ними должна осуществляться на других этапах повышения качества данных.

Другим важным аспектом формирования методологии оценки качества данных следует считать необходимость классификации проблем, связанных с качеством данных, применительно к одному из трех уровней: концептуальному, аналитическому или техническому. В то же время наиболее критичными проблемами следует считать те, которые связаны с концептуальным уровнем. Наличие таких проблем указывает на то, что стратегия сбора данных имеет серьезные недостатки, а собранные данные недостаточно отражают исследуемые бизнес-процессы. Если обнаруживается, например, что недостаточно данных для всестороннего описания предметной области, необходимо использовать методы обогащения данных для решения проблемы. Гораздо реже оказывается, что объем данных чрезмерен, т.е. часть из них не имеет отношения к изучаемой предметной области, и необходимо принять меры по уменьшению размерности исходного набора данных путем уменьшения количества признаков или количества их значений.

Такие факторы, как шумы данных, аномальные значения, противоречивые и дублирующие записи и пропуски, обусловливают проблемы качества данных, которые относят к аналитическому уровню. Необходимо учитывать, что для него характерна субъективность оценки качества данных. Так, шум обычно происходит в виде быстрых изменений значений ряда данных, мешающих выявить общие закономерности и тенденции. Но то, что даже для одного и того

же аналитика в одной ситуации будет просто шумом, в другом случае может считаться ценной информацией.

Аномалии также не так просты, поскольку может быть довольно сложно однозначно сказать, являются ли они просто ошибками оператора или отражают реальные события, исключение которых приводит к потере важной информации. Наконец, идентификация дублирующихся записей должна проводиться очень тщательно, поскольку вполне вероятно, что два клиента с одинаковыми именами и с разными адресами на самом деле являются совершенно разными фирмами, что можно проверить, дополнительно сравнив их банковские реквизиты.

К проблемам технического уровня принято относить проблемы, связанные с нарушениями структуры данных, их целостности и завершенности, неправильными форматами и кодировкой и т.д., что препятствует интеграции данных, их загрузке в HD и в аналитические системы. Такие проблемы довольно просто выявляются по формальным признакам и устраняются.

Рассмотренная классификация проблем с качеством данных также важна для того, чтобы определить место для их решения. Технический - проблемы уровня решаются только в процессе ETL, источники данных, процессы ETL и аналитические системы могут быть местом для решения проблем аналитического уровня, а проблемы концептуального уровня потребуют доработки стратегии сбора данных и/или аналитических процессов. В любом случае требуется внимание к проблемам каждого уровня, потому что, например, если есть проблемы на концептуальном уровне, то анализ накопленных данных оказывается совершенно бессмысленным, даже если они абсолютно верны. Наличие технических проблем в данных, независимо от того, какую ценную информацию содержат эти данные, просто не позволит предоставить ее аналитику, поскольку такие данные не могут быть загружены HD. Напротив, неверные с точки зрения анализа данные дойдут до аналитика, но вряд ли его обрадуют, поскольку они не могут обеспечить значимых и надежных результатов при использовании даже самых передовых аналитических технологии.

Что касается конкретных технологий оценки качества данных, то вполне естественное желание разработчиков минимизировать трудозатраты при

одновременном повышении качества. Это данных делает актуальным широкое использование так называемого профилирования данных, в ходе которого анализируется следующая информация: тип, длина, шаблон и диапазон допустимых значений - для каждого атрибута (поля). Однако, если объем исходных данных не слишком велик или среди них можно заранее определить наиболее значимую информацию, то нет необходимости оценивать качество данных. визуальными методами следует пренебречь, используя для этого как встроенные средства визуализации, так и дополнительные программные средства. Конечно, "камнем преткновения" являются трудно формализуемые ошибки, которые обнаруживаются с помощью более изощренных методов. Эти методы обычно требуют четкого знания того, какими должны быть качественные данные, что не всегда возможно определить заранее. Именно так в таких случаях, когда нет стандартных решений, требуется не только профессионализм, но и творческий подход, поиск неординарных ходов для решения очень нетривиальной задачи повышения качества данных.

Наконец, мы никогда не должны забывать о таких простых, но достаточно эффективных способах борьбы за качество данных, как наличие четких, однозначно понятных технологических инструкций по вводу данных, поощрение сотрудников, допустивших наименьшее количество ошибок, а также дублирование каналов ввода данных.

Теперь стоит рассмотреть несколько наборов данных из моей научной работы.

Примеры

Chandra Source Catalog

Chandra - каталог космических источников рентгеновского излучения.

Full	_RAJ2000 "h:m:s"	_DEJ2000 "d:m:s"	2CXO	RAICRS deg	DEICRS deg	r0 arcsec	r1 arcsec	PA deg	Ns
1	04 30 16.967676	+35 12 08.46389	J043016.9+351208	67.5706987	35.2023511	0.891	0.757	1.579e+01	1
2	04 30 16.996214	+35 16 46.66103	J043016.9+351646	67.5708176	35.2796281	0.854	0.847	1.563e+02	1
3	04 30 17.093373	+35 19 26.36620	J043017.0+351926	67.5712224	35.3239906	0.795	0.774	1.169e+02	1
4	04 30 17.160629	+35 16 15.59151	J043017.1+351615	67.5715026	35.2709976	0.741	0.734	4.120e+01	1
5	11 28 21.431778	+58 32 45.37552	J112821.4+583245	172.0892991	58.5459376	1.037	0.857	1.073e+02	1
6	04 30 17.270160	+35 15 37.77129	J043017.2+351537	67.5719590	35.2604920	0.712	0.711	9.323e+01	1
7	04 30 17.294711	+35 16 03.14780	J043017.2+351603	67.5720613	35.2675411	0.801	0.783	1.711e+02	1
8	04 30 17.613629	+35 22 38.00772	J043017.6+352238	67.5733901	35.3772244	2.083	1.587	1.184e+02	1

Представленные не все столбцы данных, всего их более 30

Попробуем оценить в общем данный датасет по представленным в начале характеристикам от 0 до 5 субъективно.

Доступность - 5: Представленный каталог находится в открытом доступе и позволяет пользоваться собой каждому желающему пользователю

Своевременность – 1: Каталог был составлен в определенный год, и не обновляется с ходом времени. Однако периодически выпускают новые каталоги, являющиеся новыми версиями старого. Это в целом свойственно большинству астрономических каталогов.

Доверие – 4: Для многих столбцов таблицы представленная ошибка измерения. Хоть и точность в некоторых местах и низкая, но мы хотя-бы знаем, насколько данным можно доверять.

Документация – 5: Для каждого столбца данных представлено описание величины. Найти это можно на специализированном сайте для хранения космических данных Vizier.

Точность – 3: В некоторых местах точность страдает, но это обусловлено недостатками аппаратуры телескопа. Для каждого отдельного обзора характерна своя точность.

Согласованность – 5: Все данные в столбцах являются элементами одного типа. Компания, предоставившая данный датасет позаботилась об этом вопросе.

Полнота – 3: Во многих важных столбцах (координат локализации, сила излучения и т.д.) полнота 100%. Однако в некоторых других существую пустые ячейки. Обусловлено это трудностью их вычисления или небольшой важностью этих показателей.

Проверяемость – 2: Перепроверить данные сложно, так как для этого требуется повторный обзор некоторого участка космоса, а в этот момент космическая картина может уже поменяться. Космические объекты движутся, хоть и не сильно

Удобство чтения – 2: Сложно считывать данные пользователю, не являющемуся экспертов в предметной области, возможно из за специфики области. Однако эксперту будет не сложно разобраться.

Структурность – 4: У данного каталога присутствует хорошая структура. Данные на сайте разделены на сегменты, описывающие определённые характеристики процесса.

Доступ.	Своевр.	Довер.	Докум.	Точн.	Соглас.	Полнот.	Провер.	Удобство.	Структ.
5	1	4	5	3	5	3	2	2	4

Вычислив среднее арифметическое, можно получить 3.4, что кажется неплохой оценкой. Однако некоторыми действиями можно улучшить показатели.

Полнота – Для заполнения пустых строк можно выбрать одну из нескольких стратегий:

1) Удалить строки с пропущенными значениями. Данный плох в данном каталоге, так как придется удалить слишком много строк.
2) Не использовать для решения задач моей научной работы столбцы с пропущенными данными. Данная стратегия имеет смысл.

3) Заполнить пропуски некоторым значением. Например, средним арифметическим для численных значений или самым частым элементом для категориальных.

Для повышения качества целесообразно использовать все 3 способа. Можно выделить важные для задачи столбцы, если в них есть пустые значения, то в зависимости от количества пропусков, стоит либо удалять строки, либо заполнять их каким-то значением.

Так же можно и вовсе не обращать внимание на пропуски, так как для использования обработанного каталога в дальнейшем пустые строки не будут сильно мешать.

От данного каталога мне нужны только первые два столбца (координаты космического объекты), которые полные. Поэтому обращать внимание на эту проблему я не буду.

Удобство чтения нивелируется ненужностью большинства столбцов этой таблицы для моей задачи.

Своевременность и проверяемость никак изменить нельзя из-за особенности области данных. По крайней мере в улучшении постобработка никак не влияет.

Таким образом, получилось улучшить некоторые низко оцененные характеристики (полнота, удобство) исходя из особенности поставленной задачи. Данный каталог получился хорошим для работы.

Рассмотрим следующий каталог

Pan-STARRS catalog

Pan-STARRS – каталог оптических объектов, состоящий из 10 млрд строк.

Оценим его по прошлым характеристикам

Доступность - 4: Представленный каталог находится в открытом доступе хотя и сложен для скачивания, так как его размер более 1 тб.

Своевременность – 1: Как и прошлый каталог, этот не обновляется со временем. Выпускаются новые версии периодически.

Доверие – 4: Для многих столбцов таблицы представленная ошибка измерения. Хоть и точность в некоторых местах и низкая, но мы хотя-бы знаем, насколько данным можно доверять.

Документация – 5: Для каждого столбца данных представлено описание величины. Найти это можно на специализированном сайте для хранения космических данных Vizier.

Точность – 3: Точность у данного каталога ниже, чем у Chandra. И также как и там указана позиционна ошибка.

Согласованность – 5: В каждом столбце присутствует только один тип.

Полнота – 3: Во многих важных столбцах (координат локализации, сила излучения и т.д.) полнота 100%. Во многих других же много пустых ячеек.

Проверяемость – 2: Перепроверить данные сложно, так как для этого требуется повторный обзор некоторого участка космоса, а в этот момент космическая картина может уже поменяться. Космические объекты движутся, хоть и не сильно

Удобство чтения – 1: Данные крайне сложно читать из-за размера каталога. В нем присутствует много различной информации, которая скорее всего не пригодится для большинства задач

Структурность – 4: У данного каталога присутствует хорошая структура. Данные на сайте разделены на сегменты, описывающие определённые характеристики процесса

Доступ.	Своевр.	Довер.	Докум.	Точн.	Соглас.	Полнот.	Провер.	Удобство.	Структ.
4	1	4	5	3	5	3	2	1	4

Средний балл получился 3.2

У данного каталога очень много похожих характеристик с предыдущим, так как каталоги в целом из одной отрасли. Но сравнив с предыдущим можно

сказать, что данный каталог немного хуже. Однако проведя аналогичные с предыдущим действия для повышения качества, можно получить хороший результат.

Таким образом, каталоги из моей научной работы получились неплохие. После предобработки они стали применимы для задач, необходимых для выполнения так, что можно и не применять другие методы повышения качества.

Вывод

Сегодня во многих отраслях используются большие данные, как в государственных, так и в коммерческих, поэтому очень важно, чтобы данные были качественными, так как от них зависит на сколько качественным будет прогноз. С увеличением количества данных увеличивается скорость получения этих данных, повышается необходимость быстрой обработки, поэтому нужно понижать размерность данных и быстро вычленять полезную информацию из них. Это можно делать с помощью методов повышения качества.

В данном реферате был проведен обзор современных методов оценки качества и характеристик больших данных. Рассмотрены несколько способов повышения качества, в том числе и на примере из моей научной работы.

Сегодня задача повышения и оценки качества данных является очень важной, но у нее все еще нет единого качественного решения. Но существует множество различных способов повышения качества для задач, характерных определенным предметным областям.

Список источников

1) The Challenges of Data Quality and Data Quality Assessment in the Big Data Era [HTML] https://datascience.codata.org/articles/10.5334/dsj-2015-002/

2) Data quality [HTML] https://sci-hub.ru/10.1016/0950-5849(90)90146-I

3) Cisco Annual Internet Report (2018–2023) White Paper [HTML] https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

4) IMPROVEMENT IN DATA QUALITY IN THE CONTEXT OF MODERN ANALYTICAL TECHNOLOGIES [HTML] https://cyberleninka.ru/article/n/povyshenie-kachestva-dannyh-v-kontekste-sovremennyh-analiticheskih-tehnologiy/viewer
5) Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets [HTML] https://arxiv.org/pdf/2108.05935.pdf

6) Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations [HTML] https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations

7) If Your Data Is Bad, Your Machine Learning Tools Are Useless [HTML] https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

8) ISO 9000 на русском [PDF] https://www.istu.edu/docs/education/fgos_14/ISO_9000-2005rus.pdf

9) What Is Data Quality and Why Is It Important? [HTML] https://www.ataccama.com/blog/what-is-data-quality-why-is-it-important

10) Data Quality: A Comprehensive Overview [HTML] https://blog.hubspot.com/website/comprehensive-overview-of-data-quality

11) Data quality [HTML] https://www.techtarget.com/searchdatamanagement/definition/data-quality

12) McGilvray, D. (2010) Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Beijing: Publishing House of Electronics Industry.

13) Silberschatz, A., Korth, H., & Sudarshan, S. (2006) Database System Concepts, Beijing: Higher Education Press.

14) Guide to Data Quality Management [HTML] https://www.scnsoft.com/blog/guide-to-data-quality-management

15) Bad Data Costs the U.S. $3 Trillion Per Year [HTML] https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

16) Data quality [HTML] https://bi-survey.com/data-quality-master-data-management

17) How to Improve Your Data Quality [HTML] https://www.ataccama.com/blog/what-is-data-quality-why-is-it-important

18) ETL (Extract, Transform, Load) [HTML] https://www.ibm.com/cloud/learn/etl

19) Data Quality Improvement [HTML] https://www.sciencedirect.com/topics/computer-science/data-quality-improvement

20) How to measure data quality [HTML] https://clearbit.com/blog/how-to-measure-data-quality

21) Data Quality [HTML] https://www.cloverdx.com/explore/data-quality

22) Evaluating Data Quality for Machine Learning Models at Scale [HTML] https://www.capitalone.com/tech/open-source/data-profiler-evaluating-data-quality-at-scale/

23) Data quality evaluation [HTML] https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/quality-qualite-eng.htm

24) Strasbourg astronomical Data Center [HTML] https://cds.u-strasbg.fr

25) Pan-STARRS1 data archive home page [HTML] https://outerspace.stsci.edu/display/PANSTARRS/

26) X-Match [HTML] https://archive.stsci.edu

27) XMM-Newton [HTML] http://xmm-catalog.irap.omp.eu