Санкт-Петербургский государственный университет

ТИМОФЕЕВ Александр Валентинович

Выпускная квалификационная работа

Адаптация фреймворка BGX для консенсуса горизонтально интегрированной среды

Направление 02.03.02
«Фундаментальная информатика и информационные технологии»
ООП СВ.5003.2016 «Программирование и информационные технологии»

Научный руководитель:
профессор кафедры КТиПА,
д.ф.-м.н. Богданов
Александр Владимирович

Рецензент: доцент кафедры
КММС, к.ф.-м.н. Корхов
Владимир Владиславович

Санкт-Петербург
2020

Содержание

Введение

Сегодня человечество живет в эпоху цифровизации, когда неразрывным спутником повседневной жизни является широкое применение технологий: многочисленные IT системы, работа пользователей в социальных сетях, сайты в Internet ежесекундно создают 10^9 байт. Это приводит к значительному росту объема данных, вариативности источников информации, скорости их изменения. Согласно отчету [1] 90% данных было создано за последние несколько лет.

Но ценность данных, возможность извлекать из них знания, принимать решения — существенно зависит от качества. Для упорядоченных, регулярных (структурированных) данных были найдены простые и эффективные решения, внедренные в инструменты управления данными. Однако, в случае больших данных (Big Data) контролировать качество значительно сложнее из-за большого разнообразия их источников, которые порождают конфликты, противоречия, потери связей и другие проблемы.

Наиболее критично вопрос о качестве данных возникает в важнейшей для цифровой экономике проблеме — построении обоснованных консенсусов.

Постановка задачи

Переход к консенсусу в горизонтально интегрированной структуре упирается в целый ряд проблем: малая скорость транзакций, трудности с глобальным арбитражем, и так далее. Но все они упираются в проблему повышения качества данных. Блокчейн обладает свойствами неизменяемости и обеспечения целостности, что исключает возможность изменения данных, записанных в блокчейн, и тем самым улучшает контроль качества. Консенсус F-BFT хорошо подходит, как для построения горизонтально интегрированной системы, так как может быть масштабирован вертикально и горизонтально, так и для контроля качества данных.

Примером контроля качества данных на базе консенсуса F-BFT могут быть написанные семейства транзакций, которые отбирают данные по заданному критерию. Поэтому для демонстрации качества контроля данных в качестве такого примера был выбран процессор транзакций, контролирующий качество данных при вставке в реестр транзакций информации по ценным бумагам.

Поставленная задача может быть разделена на следующие подзадачи:

1) изучение понятия качества данных и метрик качества;
2) изучение контроля качества данных при помощи технологии блокчейн;
3) изучение архитектуры DGT;
4) написание процессора транзакций для контроля качества данных.

Обзор литературы

Согласно [2] (Gartner Report) значительного прогресса в вопросе качества больших данных можно достичь за счет проверок в реальном времени и создания надежной среды хранения распределенных данных с помощью блокчейн технологий. В этом направлении ведется целый ряд исследований, например, [3], в которых предлагается проводить такие проверки непосредственно в реальном времени за счет валидации при вставке данных в распределенный реестр или с использованием смарт-контрактов, обеспечивающих проверку данных.

В то же время нельзя не отметить, что прямое использование классических блокчейн-систем для контроля качества данных имеет ряд ограничений:

- Скорость работы и возможность масштабирования классических блокчейн сетей ограничены энергозатратными механизмами проверки (механизмы консенсуса) и не могут обеспечить высокопроизводительную обработку данных;
- Поддержка публичных сетей за счет механизмов майнинга или подобных вероятностных методов является дорогой с точки зрения использования вычислительных мощностей на единицу обрабатываемых данных;
- Блокчейн сети часто используют для своей работы GOSSIP – протоколы, порождающие избыточный трафик, который затрудняет работу с потоковыми данными, превалирующими над традиционной пакетной обработкой.

Учитывая перечисленное выше, представляется актуальным использование решений распределенной обработки данных гибридного характера: сохраняющими основные положительные свойства блокчейна в

части проверок в реальном времени и хранения неизменной копии реестра, но уклоняющихся от отмеченных выше недостатков.

Ниже представлено решение на базе платформы DGT, представляющее собой распределенную гибридную сеть с возможностью хранения распределенного реестра в виде DAG (Directed Acyclic Graph). Платформа работает с использованием консенсуса F-BFT [4], позволяющего проводить проверки качества данных в реальном времени без падения скорости из-за конкуренции узлов.

Общий механизм предлагаемого решения дается следующим алгоритмом:

- Один из узлов (виртуальных серверов) сети пытается вставить данные в реестр. Для этого он должен валидировать информацию за счет ее проверки с другими узлами;
- Для этой цели узел отправляет данные с использованием специального механизма пермалинков в топологически соседние узлы, каждый из которых проверяет данные на предмет их соответствия правилам валидации (корректность, имеются ли уже копии таких данных в реестре, соответствуют ли данные правилам полноты и т.п.);
- Если данные прошли проверку, соседние узлы «голосуют за вставку» и данные добавляются узлом-лидером, поддерживающим группу узлов;
- В процессе проверок могут осуществляться коррекции данные, характеризующие как сам пакет (микро-пакет) данных, так и его источник. Это обстоятельство позволяет непосредственно в процессе обработки оценивать метрику качества данных, как это сделано, например, в [5].

Поскольку DGT является мульти-транзакционной сетью (одновременно допустимы разные семейства транзакций), для моделирования вопросов обработки качества в настоящей работе представлено отдельное семейство транзакций, которое имитирует вставку в реестр транзакций по ценным бумагам.

Данный пример выбран как важный случай потоковых данных, позволяющих оценить и сам источник, и качество информации по торгам (например, насколько она изменяется искусственно). Поскольку такие данные могут поступать из разных источников, само поведение систем в значительной степени контролируются торговыми алгоритмами и ботами, оценка качества данных может быть выполнена за счет статистического анализа тиковых баров (tick bars – минимальных движений цен по каждой акции).

Глава 1. Качество данных

1.1. Понятие качества данных

Качество данных — характеристика, показывающая степень пригодности данных к использованию.

В соответствии со стандартом ISO 9000:2015 основными критериями качества являются полнота, достоверность, точность, согласованность, доступность и своевременность.

Основными проблемами данных является:

- Пропущенные значения;
- Дубликаты;
- Противоречия;
- Аномальные значения и выбросы;
- Шум;
- Отсутствие полноты данных;
- Нарушение целостности данных;
- Некорректные форматы и представления данных;
- Фиктивные значения;
- Ошибки ввода данных;
- Нарушения структуры.

Независимо от того, какие факторы снижения качества присутствуют в данных, с ними необходимо бороться. Это обеспечивается в несколько этапов, самыми важными из которых являются:

- Профайлинг — исследование данных с целью выявления проблем и выработку стратегии их решения.
- Очистка — применение различных методов для разрешения обнаруженных проблем: восстановление пропущенных значений, редактирование аномалий, обработка дубликатов и противоречий и так далее.

Обеспечение качества данных неразрывно связано со всем циклом поддержки данных:

- Определение целей;
- Оценки и измерения;
- Анализ отклонений;
- Улучшение;
- Внедрение;
- Контроль.

Большие данные значительно увеличивают вероятность возникновения ошибок из-за влияния их характеристик (размера, скорости, разнообразия, точности и стоимости), которые отражают их распределенный характер:

- Унифицированная информация часто включает в себя несколько типов данных (структурированные, полуструктурированные, неструктурированные), что затрудняет сверку данных даже между данными, имеющими одинаковые значения;
- Семантические различия в определениях могут привести к тому, что одни и те же позиции заполняются по-разному.
- Различия в форматах и синтаксисах.

Для оценки качества данных используются метрики качества данных из табл.1, описанные в статье [6]:

Таблица 1. Оценки качества данных

Компонент	Нотация	Определение
Релевантность	R1	Индекс востребованности, определяемый по частоте доработки витрин
	R2	Показатель по статистике недоступности информации
Достоверность и точность	A1	Коэффициент вариативности
	A2	Коэффициент отказов процедур ETL
	A3	Показатель уточнений
	A4	Размер отказа в классификациях
	A5	Показатель приемлемости отчетов
Интегрированность	I1, I2, I3	Обработка конфликтов при сборе данных
Полнота	AC1	Частота обращения к данным
	AC2	Необходимость доработки

Согласованность	CH1	Оценки времени очистки по отношению к общему циклу загрузки
	CH2	Дистанция между начальным и конечным векторами данных
	CH3	Совпадение с результатами, полученными из других источников
Своевременность	T1, T2, T3	Длина временной шкалы и задержка данных

Индекс качества данных (Data Quality Index) равен нормированной сумме показателей по всем измерениям качества данных. Ниже приведена обобщенная формула индекса качества данных:

$$DQT = \frac{\sum \omega^i R_i \omega^j A_j \omega^k I_k \omega^l AC_l \omega^m CH_m \omega^n T_n}{\sum \omega^i \omega^j \omega^k \omega^l \omega^m \omega^n}$$

Аналитик компании SAS в своей работе [7] к вышеперечисленным метрикам качества добавляет уникальность. Это позволяет элементам встречаться один раз в наборе, что помогает избежать дублирования.

На рис. 1 изображен процесс получения метрик качества. Получение показателей качества и расчет общего индекса осуществляется с использованием внедренных агентов в процессы профайлинга и очистки.

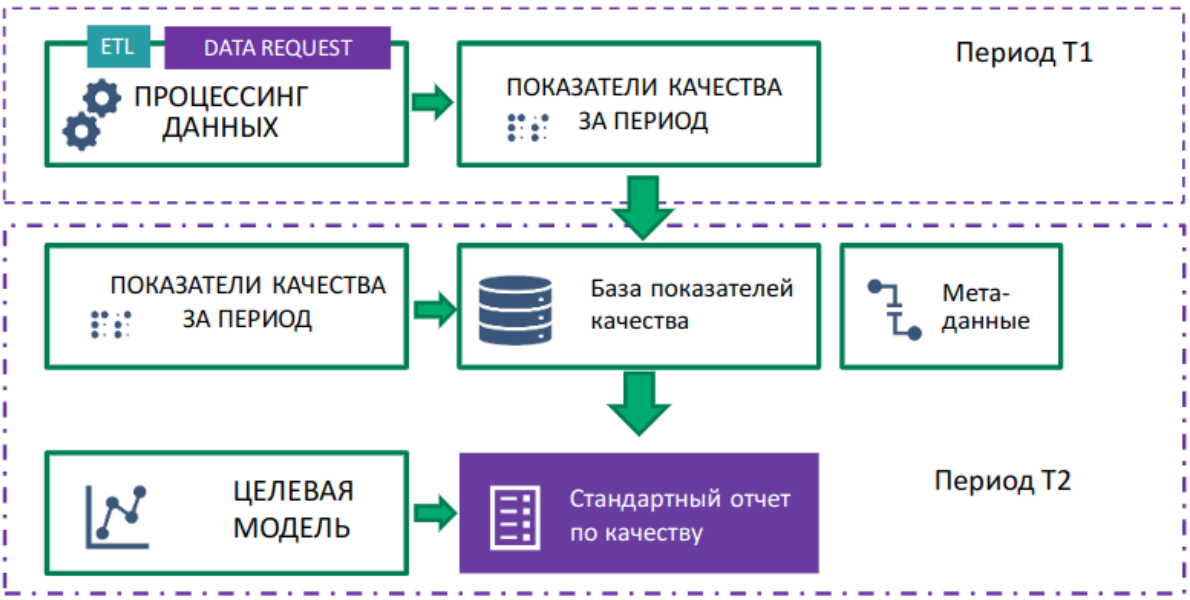

Рис 1. Алгоритм получения метрик качества [8]

Также метрики качества могут вычисляться по паттернам, например таким как:

- Использование семантических шаблонов для имен, географии и терминов;
- Обнаружение мошенничества с использованием закона Бенфорда;
- Статистика с пороговыми значениями индикаторов;
- Анализ набора столбцов;
- Расширенный анализ соответствия;
- Анализ корреляции столбцов времени.

1.2 Качество больших данных

В статье [9] авторы описывают термин большие данные тремя характеристиками — большой объем, высокая скорость и большое разнообразие. Они отмечают, что наборы данных настолько велики и сложны, что традиционных приложений становится недостаточно для сбора,

обработки, хранения и анализа данных. Авторы подчеркивают, что большие данные часто не структурированные и поэтому требуют надлежащего преобразования для улучшения качества данных.

Методы обеспечения качества данных, которые традиционно используются при работе со структурированными данными, плохо работают для больших данных. Многие правила качества данных, используемые со структурированными данными, не применяются, если данные не организованы и не управляются как реляционные таблицы.

Профилирование данных — это хороший первый шаг в оценке качества данных. Но для больших данных она отличается от структурированных данных. Структурированные методы профилирования столбцов, таблиц и кросс-таблиц не могут быть легко применены к большим данным. Инструменты виртуализации данных могут создавать представления строк или столбцов для некоторых типов больших данных, где эти представления затем могут быть профилированы с использованием реляционных методов. Этот подход обеспечивает полезную статистику содержания данных, но не дает полного представления о форме данных.

При использовании больших данных качество должно оцениваться как соответствующее поставленной цели. При использовании аналитики потребность в качестве данных может сильно варьироваться в зависимости от конкретного случая использования. Определение качества зависит от варианта использования, и каждый вариант использования имеет уникальные потребности в точности, достоверности, своевременности и полноте данных.

Оценка качества данных необходима для больших данных, но это не так просто, и она не может быть полностью субъективной. Нужна структура, которая обеспечивает критерии и руководство для оценки качества данных. Автор статьи [10] рекомендует структуру, основанную на трех измерениях:

- Характеристики данных, включающие в себя сами данные, доступные метаданные и источник данных.
- Полезность данных, включая их интерпретируемость, релевантность и точность.
- Обработка данных, включающая прием внутрь, уточнение и потребление.

Эти три измерения, каждое из которых имеет три подтемы, дают девять областей для оценки данных. Пересечение каждого измерения с другими приводит к 27 вопросам, которые могут помочь обеспечить объективность в работе по оценке качества данных.

1.3. Блокчейн и качество данных

Блокчейн — это связный список, состоящий из серии блоков. Каждый блок содержит данные и хэш-указатель на предыдущий блок в списке. Хэш-указатель — это указатель на место хранения информации и хэш от этой информации. Имея хэш указатель можно:

1. Запросить информацию, на которую он указывает;
2. Верифицировать то, что хэш не изменился, как следствие, не изменилась информация.

Таким образом, если злоумышленник попытается изменить содержимое одного из блоков, хэш этого блока изменится и не будет совпадать с сохранным ранее.

Авторы статьи [11] утверждают, что блокчейн позволяет улучшить качества принятия решений, поскольку его непрерывная проверка делает данные более надежными и заслуживающими доверия.

В статье [12] также указано, что блокчейн по своей природе обеспечивает гарантии качества в отношении данных, хранящихся в нем.

Хэши, связывающие блоки, предотвращают подделку данных, а использование криптографических подписей обеспечивает их происхождение и неотрицание.

Также авторы этой работы предлагают некоторые подходы для контроля качества в блокчейн. Контроль качества данных может осуществляться при помощи смарт-контрактов, которые обеспечивают необходимую гибкость для обеспечения различного контроля качества. Клиентские транзакции всегда обращаются к смарт-контракту, содержащему логику управления качеством данных. Транзакции, осуществляемые без смарт-контрактов, не будут вноситься в блокчейн, чтобы избежать записи некачественных данных.

Глава 2. Платформа DGT

2.1 Архитектура и особенности работы DGT

DGT является процессинговой системой по обмену информацией между независимыми узлами. Основой продукта является программное обеспечение узла, объединяющего все необходимые сервисы для взаимодействия с другими узлами, образующими сеть.

Данная сеть не является одноранговой, в которой все узлы равны друг другу. Вместо этого сеть делиться на группу узлов — кластеры. Такое представление сети позволяет строить сложные правила согласования информации между группами узлов, объединяемыми под названием F-BFT консенсус.

DGT изначально построен на фреймворке Hyperledger Sawtooth. DGT использует все низкоуровневые технические решения Sawtooth, добавив ряд особенностей таких как:

- Замена одноранговой сети на федеративную;
- F-BFT консенсус, основанный на достижении консенсуса в кластерах и затем распространении проверенных данных на всю сеть;
- Применение средства DAG, вершины которого содержат пакеты транзакций, а ребра образуются ссылками-хэшами на предыдущие транзакции пакеты.
- Создание механизма токенизации, позволяющего участникам системы выпускать собственные эквиваленты ценности для информационного обмена.

Архитектура платформы DGT представлена в табл. 2.

Таблица 2. Архитектура DGT

№	Подсистема	Описание
Валидатор		Объединение группы подсистем вокруг процесса получения, проверки и отбора данных для вставки в базу данных (реестр)
1	Networking	Подсистема представлена унаследованной библиотекой ZeroMQ и отвечает, как за взаимодействие с другими узлами, так и за взаимодействие с другими компонентами узла - процессором транзакций, движком консенсуса, REST API.
2	Encryption	Библиотека формирования hash-функций и формирования цифровой подписи.
3	State	Хранилище данных, реестр, DAG.
4	Journal	Этот модуль представляет из себя группу компонентов работающих над обработкой транзакций и их вставкой в реестр.
5	Consensus	Консенсус F-BFT. В процессе консенсуса производится проверка правил для выбранного семейства транзакций, а затем голосование.

6	ORACLE	Интеллектуальный компонент по получению дополнительных условий в отношении транзакций. Например, это может быть использовано для выявления мошеннических транзакций или отбора названий организаций с помощью нейронной сети.
Внешние компоненты узла		Набор подключаемых компонент
7	TRANSACTION PROCESSORS	Модуль, реализованный в виде отдельного сервиса, реализующий поддержку тех или иных семейств транзакций. Процессоры транзакций ответственны за проверку транзакций и производимые действия.
8	REST API	Компонент, обеспечивающий взаимодействие клиентов с ядром узла посредством HTTP/JSON. Обработка ведется через ZMQ/Protobuff.
Клиенты		
9	CLI	Интерфейс командной строки, позволяющий обрабатывать информацию внутри узла через стандартизованный API. Является основным средством администрирования узла.
10	DASHBOARD	Облегченный Web-портал с визуализацией основных параметров сети.
11	Mobile App	Инструмент управления цифровым контентом.

Согласованность транзакций образует F-BFT консенсус, в рамках которого узлы при вставке транзакций «голосуют» за транзакции — сначала внутри кластера, затем вовне.

«Голосование» представляет собой проверки транзакции через набор правил, разделяемых на два типа:

- Верификация — проверка корректности транзакции.
- Валидация — проверка более сложных правил, к которым относится удостоверение корректности цифровой подписи предыдущих голосований, осуществимость транзакции, отсутствие подозрительных признаков и т.п. Правила валидации могут быть настроены для выбранного семейства транзакций и осуществляются процессором транзакций.

F-BFT консенсус — механизм согласования, вставки транзакций в DAG (Направленный ациклический граф) и дальнейшая синхронизация обновленной информации внутри всей сети с учетом федеративной структуры узлов (голосование внутри кластеров и дальнейшее распространение по всей сети). DAG структура играет определяющую роль, поскольку позволяет распараллелить процесс вставки и проверки в разных кластерах.

В случае наличия нескольких узлов с точки зрения прохождения могут быть определены следующие роли:

- Инициатор — узел, инициирующий транзакцию (например, через взаимодействующий с ним клиент).
- Лидер — узел, собирающий транзакции внутри кластера («подсчитывающий» голосования узлов кластера).

- Арбитр — узел (узлы) за пределами кластера, проверяющий голосование внутри кластера и добавляющий транзакцию в реестр.

Основной алгоритм консенсуса выглядит следующим образом:

1. Инициатор, узел на который со стороны клиента пришла транзакция, проводит верификацию транзакции, затем вбрасывает эту транзакцию в кластер через текущего Лидера — родительского узла кластера (может быть динамическим).
2. Лидер рассылает транзакцию узлам в кластере и ждет ответов.
3. Голосование подразумевает, что каждый узел, получивший транзакцию, осуществляет серию проверок (правила валидации), затем подписывает ее или помечает отказ подписать.
4. Лидер подсчитывает количество голосований, проверяется корректность подписей при достижении необходимого числа — выводит ее за пределы кластера — на кольцо арбитров, которые представляют каждый из кластеров топологии.
5. После нескольких раундов голосований лидер может быть изменен в зависимости от SLA (показателей качества работы сервера, таких как быстродействие, время отклика и т.п.) и вероятностного выбора из числа узлов кластера.
6. Арбитр — произвольный узел за пределами кластера, проверяющий правила голосования и подписи голосовавших узлов. При корректности таких голосований арбитр вставляет транзакцию в DAG родного кластера, отсылая положительный или отрицательный результат лидеру.
7. Лидер возвращает ответ Инициатору, таким образом, транзакция считается принятой. Узлы обмениваются между собой копиями DAG

(инкрементальными наборами), далее DAG синхронизируется вдоль всей сети.

2.2 Методика учета качества поверх DGT

Платформа DGT предлагает свой подход для сохранения качества данных, описанный в статье [13]. В рамках этого подхода используются технологии, которые обеспечивают высокую скорость принятия решений и снижают потери из-за несоответствия данных:

- Интеграционный уровень системы построен на высокопроизводительном ядре DGT, которое обеспечивает формирование единого реестра основных данных и его распределение между участниками обмена информацией с большой степенью горизонтальной масштабируемости и возможностью отслеживания всей истории;
- Интеллектуальные модули, которые отслеживают данные в режиме реального времени и участвуют в построении выверенных наборов данных при одновременном измерении показателей качества;
- Разработанный API, который может подключаться не только к различным корпоративным системам и аналитическим инструментам, но также к различным инструментам управления данными и профилирования.

Ниже представлена модель для работы с большими данными на основе консенсуса F-BFT. Особенности этой модели включают:

- Обработка данных осуществляется в гибридной сети на основе консорциума, построенной по федеративному принципу: узлы сгруппированы в кластеры с меняющимися лидерами, а доступ к сети ограничен набором условий;

- Запись в реестре выполняется в результате «голосования» в кластере и последующих «одобрений» арбитражного узла. И «голосование», и «утверждение» представляют собой серию проверок-валидаций в форме расчетов с двоичными результатами;
- Каждый сетевой узел получает информацию и идентифицирует информационные объекты как один из классов основных данных;
- Если объект новый, то предпринимается попытка инициировать специализированную транзакцию для вставки данных в соответствующий реестр через механизм голосования промежуточных узлов (процесс проверки). Если новый объект одобрен другими узлами, объект добавляется в реестр, и информация распространяется в сети. Если новый объект отклоняется, узел инициализации получает ссылку на существующий объект;
- Распределенная система хранения данных (реестр) принимает форму графической базы данных (DAG, Directed Acyclic Graph), которая позволяет сосуществовать нескольким семействам транзакций для разных классов объектов, сохраняя при этом горизонтальную масштабируемость сети.

Этот подход позволяет отделить данные быстрой потоковой передачи от пакетной обработки распределенного реестра. При этом атрибуты качества рассчитываются с точки зрения целостности информации как меры конфликта для данного объекта. Эта технология полезна для очистки данных в режиме реального времени без ограничения доступности данных.

Также сложной проблемой контроля качества данных является низкая эффективность ручных проверок с увеличением объема и изменчивости данных. В таких случаях модули машинного обучения могут помочь оценить качество на ранней стадии обработки, диагностировать проблемы отсутствия

данных, наличие непредвиденных типов данных, нестандартные значения параметров, противоречия между различными наборами и т. д.

Использование искусственного интеллекта (ИИ) позволяет решать несколько важных задач:

- Очистка текстовых данных с использованием технологий Natural Language Processing (NLP) и извлечение основной информации из свободно структурированных текстов. Модули NLP могут определять степень соответствия между объектами на основе контекста;
- Обеспечение соответствия установленным стандартам и практикам управления основными данными; преобразование основных данных в стандартную форму;
- Высокоскоростное сравнение наборов данных (Entity Resolution) на основе метрик близости;
- Измерение качества данных непосредственно на основе алгоритма машины опорных векторов (SVM).

Механизм ИИ встроен в существующие механизмы проверки и работает в распределенном подходе с использованием интеллектуальных агентов (оракулов), которые непосредственно участвуют в проверке данных.

Подход, сформулированный выше, позволяет количественно оценить качество данных в режиме реального времени со следующими допущениями:

- Общая оценка качества проводится на основе взвешенного показателя, оцениваемого как количество операций, необходимых для исправления выявленной ошибки;

- Некоторые ошибки могут быть выявлены на этапе проверки данных, а другие — только во время последующего анализа. Поэтому атрибут качества для данного набора данных постоянно пересчитывается;
- Вес атрибута зависит от текущего значения надежности источника (в описываемой структуре, которая также влияет на количество проверок — «голосов»), а также от серьезности и приоритета ошибки в соответствии с соответствующим правилом проверки;
- Идентификация объекта на основе нечеткой логики и результатов нейронной сети;
- выявление аномалий и корреляция с более ранними данными. При сравнении информационных механизмов правила представляют собой картриджи (смарт-контракты), которые неотделимы от реестра.
- Признаки качества, которые неисчислимы или невозможны для оценки, отмечаются для последующего анализа.

Общий коэффициент качества можно рассчитать как средневзвешенное значение по следующим показателям:

- Количество неопознанных (неопознанных) объектов, которые были обнаружены в будущем;
- Статистика недоступности данных на основе частоты запросов;
- Обработка конфликтов сбора данных, включая аномалии и выход за пределы диапазона проверки данных;
- Расстояние между начальным и конечным векторами данных;
- Совпадение с результатами из других источников;
- Длина шкалы времени и задержка данных;
- Оценки времени очистки относительно общего цикла загрузки.

Глава 3. Построение семейства транзакций

3.1. Пример реализации семейства транзакций

В качестве примера работы DGT с качеством данных, был реализован процессор транзакций [14]. Работа этого процессора транзакций заключается в том, что пользователь по команде проверяет качество потока данных. Каждый раз при проверке новых данных они сравниваются с предыдущими данными, записанными в блокчейн. Тем самым, мы можем отклонять или принимать данные, если нас устроит их качество. В качестве данных были выбраны движения цен криптовалют на крипто валютных биржах. Такой выбор обусловлен тем, что эти данные наблюдаются в виде потока и их легко получить. В данной работе используется API [15] для получения данных с бирж. Процессор транзакций имеет 3 команды : “create”, “check”, “delete”.

“create” на вход получает имя функции контроля качества. Если такую функцию уже создал другой пользователь, то процессор выведет ошибку “Invalid action: Quality already exists:”, иначе создастся функция качества и заполнятся начальные значения тик бара такие как: время, цена первой сделки, максимальная цена всех сделок на тик баре, минимальная цена всех сделок на тик баре, цена последней сделки, сумма всех обмененных активов. Все эти данные собираются при помощи API.

“check” получает на вход имя функции контроля качества. Если такой функции не существует, то процессор выдает ошибку “Invalid action: Take requires an existing quality”, иначе процессор получает новые данные при помощи API. Далее процессор проверяет корреляцию предыдущих данных и полученных. Если качество данных приемлемое, то новые данные записываются, иначе процессор выдает предупреждение “Data of poor quality, recording denied”. Также из-за ограниченности данных, после того как данные

закончатся при попытке получить новые будет выдана ошибка “No data available”.

“delete” на вход получает имя функции контроля качества. Если такой функции не существует, то процессор выдает ошибку “Invalid action: quality does not exist”. Стоит отметить, что удаляется только функция в процессоре, все действия, которые были выполнены, будут записаны в блокчейн, и информация по тик барам останется.

При помощи этого процессора транзакций мы можем контролировать качество потока данных по биржевым торгам. Это позволяет использовать эти данные, будучи уверенными в их качестве.

3.2. Результаты работы

В работе [16] автор описывает пример контроля качества данных при помощи подсчета корреляции тик баров. В таком подходе, есть один недостаток. Злоумышленник может изменить данные предыдущих тик баров, тем самым изменить корреляцию и повлиять на контроль качества данных. Представленный мной процессор транзакций решает эту проблему, записывая все предыдущие тик бары в блокчейн, который затем распределяется по всем компьютерам сети. Таким образом, невозможна подмена исторических данных. В своей статье автор считает корреляцию Пирсона. Из рис. 2 видно, что тик бары имеют меньшую автокорреляцию по сравнению с временными шкалами.

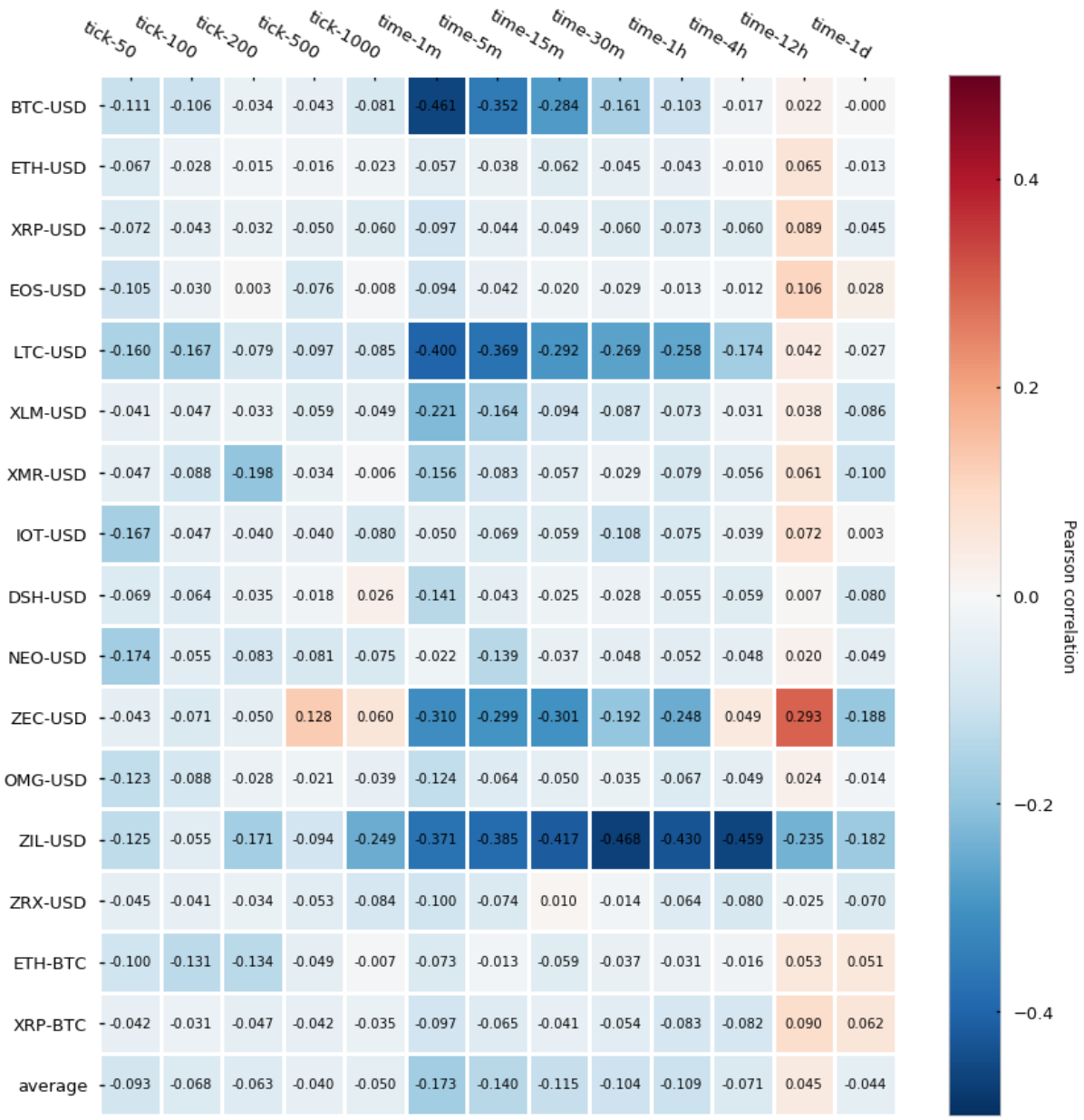

Рис 2. Корреляция Пирсона [16]

Также дополнительно был проведен тест Дарбина-Уотсона на наличие автокорреляции. На рис. 3 представлены результаты этого теста. Чем ближе значение к 2, тем меньше корреляция. Из рис. 3 видно, что тик бары имеют меньше корреляцию по сравнению с временными шкалами, что согласуется с предыдущими результатами.

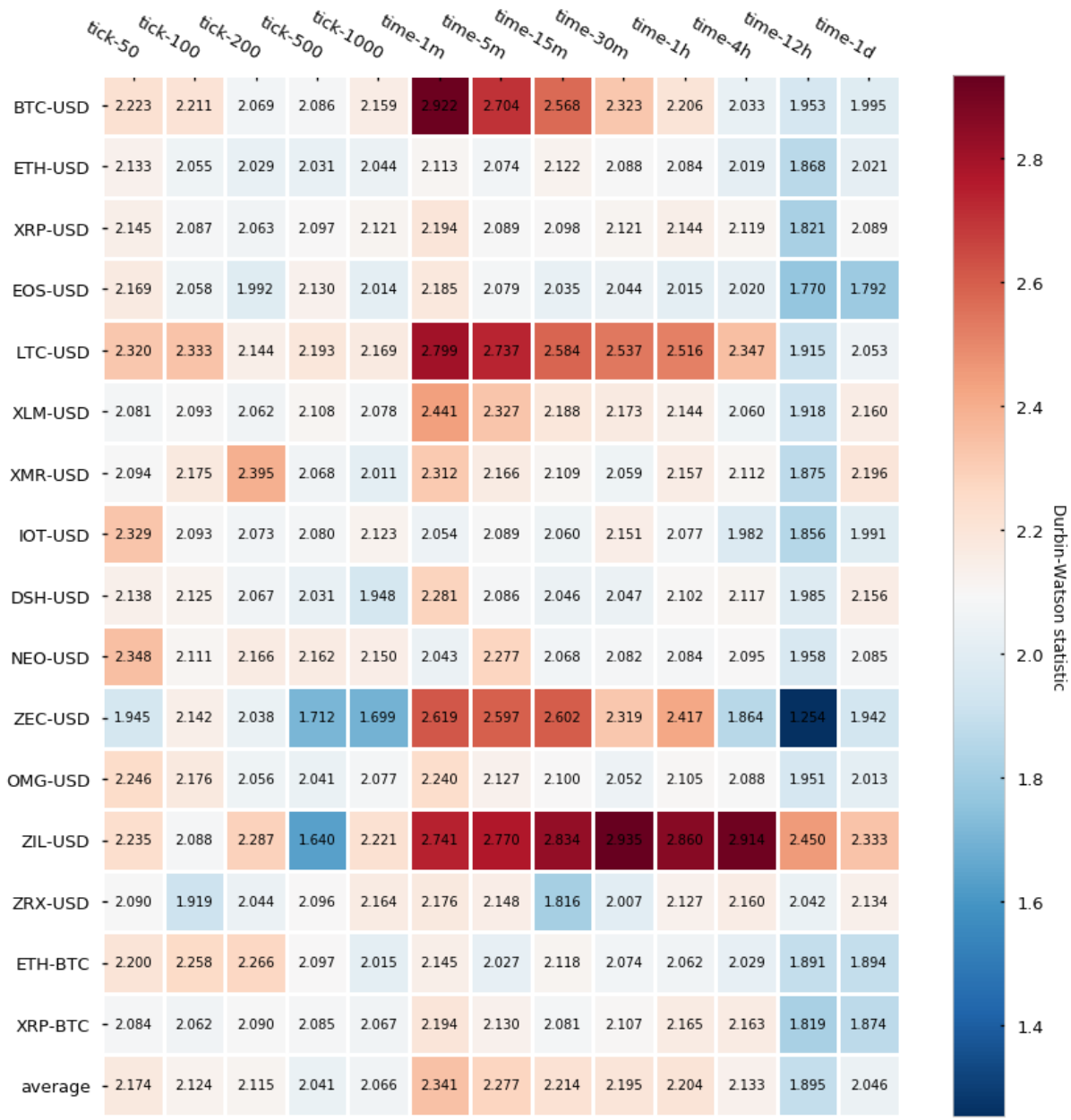

Рис 3. Тест Дарбина-Уотсона [16]

Таким образом, тик бары по сравнению с временными шкалами показывают меньшую корреляцию. Вследствие этого использование тик баров в блокчейне для сохранения качества данных предпочтительнее. Так как с понижением качества будет наблюдаться корреляция.

Как было написано выше, написанный мной процессор транзакций улучшает эти достижения, благодаря таким свойствам блокчейна как неизменяемость и обеспечение целостности.

Выводы

В ходе работы были изучены метрики качества данных и способы контроля качества данных. Стало понятно, что технология блокчейн имеет преимущества за счет своей неизменности и обеспечения целостности. В качестве реализации примера контроля качества данных была выбрана платформа DGT, так как консенсус F-BFT, используемый там, хорошо масштабируется, а также может быть использован в открытых и закрытых сетях. Процессор транзакций может работать с любыми потоковыми данными для контроля их качества. Мой выбор был остановлен на биржевых торгах, так как эти данные являются потоковыми и их легко получить. Реализованное приложение контролирует качество данных, основываясь на прошлых транзакциях.

Заключение

Контроль качества данных является одной из важнейших задач. На основе данных принимаются решения, и качество данных играет огромную роль в получении нужных результатов. В данной работе представлен пример использования DGT для контроля качества данных на базе консенсуса F-BFT. Изучены преимущества использования блокчейн для контроля данных и архитектура DGT.

Список источников

[1]. IBM Marketing Cloud, “10 Key Marketing Trends For 2017” [Электронный ресурс]: URL: https://totallygaming.com/eventblog/ice-live/ibm-marketing-experts-predict-10-key-marketing-trends-2017 (дата обращения: 30.05.2020).

[2]. Predicts 2020: Data and Analytics Strategies — Invest, Influence and Impact, Gartner Report, 2019 [Электронный ресурс]: URL: https://www.gartner.com/en/newsroom/press-releases/2020-01-30-gartner-predicts-that-organizations-using-blockchain- （дата обращения 30.05.2020）.

[3]. Hao Dai, H Patrick Young, Thomas JS Durant TrialChain: A Blockchain-Based Platform to Validate Data Integrity in Large, Biomedical Research Studies // arXiv:1807.03662. 2018.

[4]. Bogdanov A., Uteshev A., Khvatov V. Error Detection in the Decentralized Voting Protocol // Computational Science and Its Applications – ICCSA 2019. ICCSA 2019. 2019. LNCS 11620, Springer. P. 485-494.

[5]. Courtney Napoles, Keisuke Sakaguchi, Matt Post, Joel Tetreault. Ground Truth for Grammatical Error Correction Metrics, Johns // Hopkins University. 2016.

[6]. List of Conformed Dimensions of Data Quality, CDDQ Open Standard [Электронный ресурс]: URL: http://dimensionsofdataquality.com/alldimensions (дата обращения 30.05.2020).

[7]. John Bauman Data quality management: What you need to know. [Электронный ресурс]: URL: https://www.sas.com/en_us/insights/articles/data-management/data-quality-management-what-you-need-to-know.html (дата обращения 30.05.2020).

[8]. Bogdanov A., Degtyarev A., Shchegoleva N., Khvatov V. Data Quality in Decentralized Environment.

[9]. Хаммер К., Костроч Д., Кирос Г. и сотрудники Департамента статистики Большие данные: потенциал, проблемы и применение в статистике // Записка для обсуждения // МВФ. 2017.

[10]. Wells D. A Data Quality Framework for Big Data [Электронный ресурс]: URL: https://www.eckerson.com/articles/a-data-quality-framework-for-big-data (дата обращения 30.05.2020).

[11]. Predicts 2020: Data and Analytics Strategies — Invest, Influence and Impact, Gartner Report, 2019 [Электронный ресурс]: URL: https://www.gartner.com/en/newsroom/press-releases/2020-01-30-gartner-predicts-that-organizations-using-blockchain- （дата обращения 30.05.2020）.

[12]. Cappiello C., Comuzzi M., Daniel F., Meroni G. Data Quality Control in Blockchain Applications // Di Ciccio C. et al. (eds) Business Process Management: Blockchain and Central and Eastern Europe Forum. BPM 2019. Lecture Notes in Business Information Processing. Vol 361. P.166—181.

[13]. Bogdanov A., Degtyarev A., Shchegoleva N., Khvatov V. Data Quality in Decentralized Environment.

[14]. Исходный код процессора транзакций [Электронный ресурс]: URL: https://github.com/AlTimofeevM/DataQuality (дата обращения 30.05.2020).

[15]. CryptoDatum.io API documentation [Электронный ресурс]: URL:

https://documenter.getpostman.com/view/7244886/S1ENzJzL?version=latest (дата обращения 30.05.2020).

[16]. Gerard Martínez Advanced candlesticks for machine learning (i): tick bars [Электронный ресурс]: URL: https://towardsdatascience.com/advanced-

candlesticks-for-machine-learning-i-tick-bars-a8b93728b4c5 (дата обращения 30.05.2020).