

Crypto Project. Сборщик новостей. Техническое задание

1. Общая информация

Есть два новостных сайта cryptocoinsnews.com и cointelegraph.com. Наша задача выгрузить все имеющиеся новости с этих сайтов, затем с некоторой периодичностью проверять появление новой информации и тоже ее выгружать.

Загруженную новость мы должны спамить с нашей моделью данных, проверить на дубликаты и сохранить в нашем хранилище.

Проект должен быть написан на C# 6.0+, .Net Framework 4.6+

2. Структура новостных сайтов

На указанных выше сайтах информации больше чем нам нужно, поэтому в этом разделе будет указано что именно и откуда мы будем качать.

Cryptocoinsnews

Новость на этом ресурсе помечается одним или несколькими тегами. После этого новость будет доступна в той категории, тег которой был задан. Если будет задано несколько тегов - новость будет отображаться в нескольких категориях.

Есть тег News он представляет глобальную категорию для всех новостей. Проблема в том что не все новости помечаются этим тегом. Поэтому выгружать новости необходимо из каждой категории в отдельности исключая при этом повторы, которых будет не мало.

Список категорий и ссылок:

News	https://www.cryptocoinsnews.com/news/
Bitcoin Politics	https://www.cryptocoinsnews.com/bitcoin-politics/
Bitcoin Mining	https://www.cryptocoinsnews.com/bitcoin-mining/
Bitcoin Analysis	https://www.cryptocoinsnews.com/bitcoin-analysis/

Bitcoin Price News	https://www.cryptocoinsnews.com/bitcoin-price-news/
Altcoin News	https://www.cryptocoinsnews.com/altcoin-news/
ICO	https://www.cryptocoinsnews.com/ico/
Blockchain News	https://www.cryptocoinsnews.com/blockchain-news/
Bitcoin Opinion	https://www.cryptocoinsnews.com/bitcoin-opinion/
Bitcoin Technology	https://www.cryptocoinsnews.com/bitcoin-technology/
FinTech News	https://www.cryptocoinsnews.com/fintech/
Bitcoin Regulation	https://www.cryptocoinsnews.com/bitcoin-regulation/
Bitcoin Progress	https://www.cryptocoinsnews.com/bitcoin-progress/
Bitcoin Exchange	https://www.cryptocoinsnews.com/bitcoin-exchange-news/
Hacked	https://www.cryptocoinsnews.com/hacked-2/
Ethereum News	https://www.cryptocoinsnews.com/ethereum-news/
Bitcoin Education	https://www.cryptocoinsnews.com/bitcoin-education/
Banking	https://www.cryptocoinsnews.com/banking/
Sponsored Stories	https://www.cryptocoinsnews.com/sponsored-stories/
Bitcoin Business	https://www.cryptocoinsnews.com/bitcoin-business/
Bitcoin Announcements	https://www.cryptocoinsnews.com/bitcoin-announcements/
Altcoin Mining	https://www.cryptocoinsnews.com/cryptocur

	rency-mining/
Guest Contributor	https://www.cryptocoinsnews.com/guest-contributor/

Список категорий не полный. Вам необходимо самостоятельно его дополнить.

Должна быть возможность задать в конфигурации список категорий, которые необходимо исключить (не загружать). Пример: есть категория Sponsored Stories, которая явно не нужна.

Нужно обратить внимание, что новости в категории разбиты на несколько страниц. К примеру: в категории News 606 страниц. Так же и в других категориях. Необходимо выгружать новости со всех страниц в каждой категории.

Пример ссылки на вторую страницу категории News:

<https://www.cryptocoinsnews.com/news/page/2/>


Новости в категории



Название категории

Category: News

Теги/Категории




Заголовок

Bitcoin Politics / News
Worse Than 'the Russians': Kansas Panel Prohibits Bitcoin Campaign Contributions

Автор

Josiah Wilmoth / 27/10/2017

Дата



Blockchain News / ICO / News
Overstock Shares Surge 30% After \$500 Million ICO Announcement

Josiah Wilmoth / 27/10/2017

Страница новости



Категории

Автор

Дата

Заголовок

Advertisement



Get Trading Recommendations and Read Analysis on Hacked.com for just \$39 per month.

Тело
Новости

A Kansas state government commission has ruled that candidates running for office in state and local elections will be prohibited from accepting bitcoin campaign contributions.

As reported by the [Lawrence Journal-World](#), a local media outlet, the Kansas Governmental Ethics Commission has decided to bar Kansas politicians from accepting bitcoin donations when running for public office in state or local elections, citing concerns about bitcoin's pseudonymity.

Mark Skoglund, the commission's executive director, stated that the ruling was prompted by a candidate who inquired about the legality of accepting bitcoin donations.

The Federal Election Commission allows candidates running for national office to accept bitcoin campaign contributions and even stated that campaigns could invest in bitcoin under limited circumstances. A variety of candidates have taken advantage of this opportunity, most recently [Austin Petersen](#), a Republican who is running to represent Missouri in the U.S. Senate.

Petersen told CCN that cryptocurrency "represents the future of American creativity and American liberty," which is why his campaign accepts bitcoin donations:

“The greatest problem would be the strong probability of the influencing of local elections by totally unidentifiable lobbyists trying to come in,” he said. “If you think the Russians affected the presidential elections, just wait. This is what’s going to happen.”

Hellmer added that bitcoin's opacity is antithetical to the "transparency" that elections commissions are supposed to ensure. "It's totally contrary to the transparency we're asking for our political system to provide to the public," he concluded.

Featured image from Shutterstock.

Advertisement:



Posted in: Bitcoin Politics, News Tagged in: campaign donation, election, kansas

Категории

Описание автора

Posted by Josiah Wilmoth

Josiah is a former ancient and medieval literature teacher. He has been writing about cryptocurrency since 2014, and his work has been cited in Business Insider, NPR, and Yahoo! Finance. He lives in rural North Carolina with his wife and son. Email him directly at [josiah.wilmoth\(at\)cryptocoinsnews.com](mailto:josiah.wilmoth(at)cryptocoinsnews.com).

All Posts Website

Какую информацию собираем с cryptocoinsnews.com?

- Ссылка на страницу новости
- Заголовок.
- Список всех категорий
- Автор
- Дата. Внимание, здесь дата представлена только днем (27/10/2017) без указания времени. Выгружая историю новостей мы берем только дату, время ставим 00:00. При загрузке обновлений ставить то время, когда новость была обнаружена программой (только при условии того же дня). К примеру, мы выкачиваем новости сегодня (28 октября), видим что поступила новости за 28 октября, ставим для новости текущее время (таймзона EST).
- Тело новости

Cointelegraph

Как и у [cryptocoinsnews](https://cryptocoinsnews.com), здесь новости помечаются тегами и распределяются по категориям. Необходимо выгрузить все новости со всех категорий.

Список категорий и ссылок:

Blockchain News	https://cointelegraph.com/tags/blockchain
Banks News	https://cointelegraph.com/tags/banks
Bitcoin News	https://cointelegraph.com/tags/bitcoin
Bitcoin Price News	https://cointelegraph.com/tags/bitcoin-price
Tradings	https://cointelegraph.com/tags/tradings
Bitcoin Exchanges News	https://cointelegraph.com/tags/bitcoin-exchanges
Bitcoin Wallet News	https://cointelegraph.com/tags/bitcoin-wallet
Bitcoin News	https://cointelegraph.com/tags/bitcoin
Bitcoin Mining News	https://cointelegraph.com/tags/bitcoin-mining

Список категорий не полный. Вам необходимо самостоятельно его дополнить.

Должна быть возможность задать в конфигурации список категорий, которые необходимо исключить (не загружать).

Нужно обратить внимание, что новости в категории отображаются не все, их нужно подкачивать нажимая на кнопку "Load more articles".

Новости в категории

Bitcoin News ← Категория



Bitcoin is an open-source, peer-to-peer, digital decentralized currency. Using Blockchain technology, its defining characteristic is its lack of a governing authority, such as a central bank or a monetary authority. Transactions and circulation are ensured by regular users via a process of consensus. It is present anywhere, anytime, (almost) for free, and with no government/bank-imposed restrictions.

Related links:

- [Bitcoin Price](#)
- [Bitcoin Price Index](#)
- [Bitcoin 101](#)

Recent Top News Commented

Заголовок

Дата

Автор

Превью



Billionaire Investor Warren Buffett Says Leading Cryptocurrency Bitcoin in 'Bubble' Territory

14 HOURS AGO | Lisa Froelings

Billionaire investor Warren Buffett claims that the market for the leading virtual currency Bitcoin is already in bubble territory.

👁 -



Apple's Market Cap in Bitcoin's Sights: Ronnie Moas

19 HOURS AGO | Darryn Pollock

Stock Picker Ronnie Moas has another prediction, that Bitcoin will overtake Apple in five years.

👁 -

By Lisa Froelings **← Автор** **Дата** → 15 HOURS AGO

Billionaire Investor Warren Buffett Says Leading Cryptocurrency Bitcoin in 'Bubble' Territory

20996 Total views 413 Total shares

Заголовок ↗



Billionaire investor Warren Buffett has claimed that the market for the leading virtual currency Bitcoin is already in bubble territory. He also issued a criticism of the proposals for applying a value to the cryptocurrency.

Based on a [report](#) by MarketWatch, Buffett presented his views on Bitcoin and the cryptocurrency market during an annual question-and-answer session in Omaha, Nebraska in early October 2017. During his remarks, Buffett claimed that Bitcoin is a "real bubble."

"People get excited from big price movements, and Wall Street accommodates. You can't value Bitcoin because it's not a value-producing asset."

Other opinions on Bitcoin and the cryptocurrency market

Aside from Buffett, other virtual currency market observers have also issued their opinions on the latest developments in the market. In his [comment](#), Prince Al-Waleed bin Talal of Saudi Arabia has claimed that he does not believe in Bitcoin completely and expects the digital currency to fail.

"It doesn't make sense. This thing is not regulated. It's not under control. It's not under the supervision [of] any federal – elect – United States Federal Reserve or any other central bank. I don't believe in this whole thing at all. I think it's going to implode."

However, in his new blog post, New York University's "Dean of Valuation," Aswath Damodaran, asserted that Bitcoin is a true currency and not a fraud.

Follow us on Youtube

Тело ↑

Bitcoin News

Investments

Cryptocurrencies

Bubble

Warren Buffett

Markets

← Категории

Какую информацию собираем с cointelegraph.com?

- Ссылка на страницу новости
- Заголовок.
- Список всех категорий
- Автор
- Дата. Внимание, здесь дата бывает указана в нескольких форматах: “15 HOURS AGO” или “ОCT 27, 2017”.

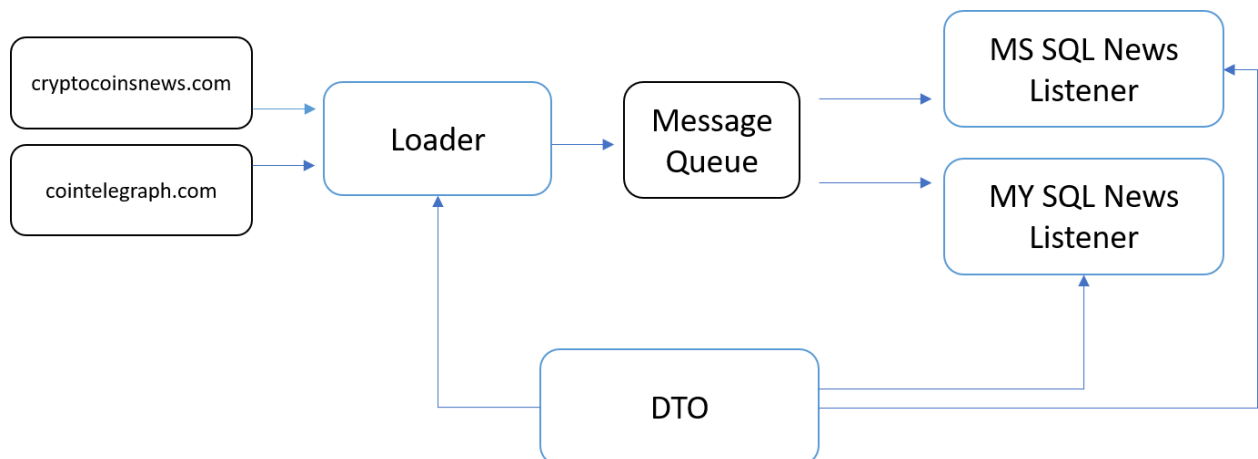
В первом случае необходимо от текущего времени EST отнять указанное кол-во часов (минут и т д).

Во втором случае, выгружая историю новостей мы берем только дату, время ставим 00:00. При загрузке обновлений ставить то время, когда новость была обнаружена программой (только при условии того же дня). К примеру, мы выкачиваем новости сегодня (28 октября), видим что поступила новости за 28 октября, ставим для новости текущее время (таймзона EST).

- Тело новости

3. Структура проекта

Мы видим реализацию проекта с использованием микросервисной архитектуры. Передача данных между компонентами системы должно быть реализовано с помощью очереди сообщений (мы используем RabbitMq).



Задача реализовать компоненты обозначенные на схеме выше синим цветом.

Loader

Это windows service hosted as console app (используется Topshelf <http://topshelf-project.com/>), которая при первом запуске выгружает с указанных сайтов историю новостей (должна быть настройка в конфигурации: грузить ли историю).

Загрузив историю, программа должна с заданной периодичностью (настройка в конфигурации) выгружать новые данные (новости).

История новостей загружаться пачками заданного объема. Объем пачки должен задаваться в конфигурации. К примеру, 100 шт. Это значит, что скачивая историю мы будем вытаскивать 100 новостей, отдавать их в очередь, а затем закачивать следующие 100.

Под выгрузкой новых данных (новостей) понимается, что программа будет выгружать только последние X новостей (X задается в конфигурации, к примеру 10) из каждой категории и передавать их в очередь сообщений.

Проблема дубликатов

Принимая во внимание сказанное выше у нас будет проблема с дубликатами, как в нутри одного новостного сайта, так и дубликатами обоих сайтах.

Было бы логичным ориентироваться на время новости, но это невозможно поскольку используемые новостные сайты не предоставляют точное время выхода новости, только день.

Мы хотим решить проблему дубликатов используя ссылки на страницы с новостями. В базе не должно быть двух записей ссылающихся на одну и ту же страницу с новостью. Мы реализуем эту проверку самостоятельно на уровне базы данных.

Это решение, к сожалению, не поможет нам в том случае, когда одна и та же новость размещена сразу на двух сайтах.

Мы готовы рассмотреть ваши предложения по решению проблемы дубликатов.

MS SQL News Listener и My SQL New Listener

Это windows service hosted as console app (используется Topshelf <http://topshelf-project.com/>). Нам нужно реализовать две службы, которые будут принимать новости из очереди и добавлять их в соответствующую базу данных.

Добавлять информацию в базу нужно не напрямую в таблицу, а с использованием хранимых процедур. Сигнатуры будут предоставлены позднее.

DTO и очередь сообщений

В качестве очереди сообщений мы используем RabbitMq. Для упрощения

взаимодействия с RabbitMq мы используем библиотеку EasyNetQ (<http://easynetq.com/>). С ее помощью мы можем передавать типы Net от поставщика к потребителю.

Чтобы это заработало нам необходимо поместить DTO типы в отдельную сборку и использовать ее во всех проектах Loader и MS SQL News Listener, My SQL New Listener.

В нашем случае в сборке должен быть определен только один тип News (подробнее дальше).

4. Наша модель данных

```
enum ProviderType
{
    cryptocurrenciesnews,
    cointelegraph
}

interface News
{
    Guid ID { get; set; }
    DateTime CreatedTime { get; set; }
    ProviderType Provider { get; set; }
    string Link { get; set; }
    DateTime Date { get; set; }
    IEnumerable<string> Categories { get; set; }
    string Author { get; set; }
    string Header { get; set; }
    string Body { get; set; }
}
```

5. Параметры конфигурации

Просим вас вынести в конфигурацию любые настройки, которые можно было бы изменить и которые могут повлиять на работу программы.

Как минимум для проекта Loader:

- грузить ли историю (по умолчанию false);
- название категорий, через запятую, которые не нужно загружать (по умолчанию пусто);
- настройки Topshelf;
- частота загрузки новых данных (по умолчанию 1 мин)

- размер пачки исторических данных (по умолчанию 100)
- размер пачки новых данных (по умолчанию 10)
- строка подключения EasyNetQ

Как минимум для проектов MS SQL News Listener, My SQL New Listener:

- настройки Topshelf;
- строка подключения EasyNetQ;
- строка подключения к базе данных

6. Тестовое окружение

После начала работы над проектом мы подготовим тестовое окружение (базы данных, очередь сообщений) и дадим вам доступ для тестирования.

7. Языки и технологии

C# 6.0+
.Net Framework 4.6+
Visual Studio 2015

Пространство имен по умолчанию во всех проектах: cryptofund

Для создания windows service использовать topshelf. Настройки topshelf выводятся в конфигурацию.

Для логирования использовать .NLog. Конфигурацию логгера поместить в NLog.config

Очередь сообщений: RabbitMq

Библиотека для работы с очередью сообщений: <http://easynetq.com/>

Базы данных: MS SQL и MY SQL

8. Требования к исходному коду

Программа должна быть составлена с использованием принципов ООП и современных практик в разработке программного обеспечения.

Код должен быть аккуратно оформлен, хорошо структурирован, иметь понятное наименование и предназначение.

Настоятельно просим вас принять во внимание, что с составленной вами программой в дальнейшем, достаточно продолжительное время, будут работать другие люди.

9. Что должно быть в результате?

После завершения работы над проектом вы должны предоставить нам исходные коды проектов (Loader, MS SQL News Listener, My SQL New Listener и DTO library) и сопутствующие материалы, если таковые будут.

Нам потребуется некоторое время на развертывание и тестирование (1 - 2 дня). Вы обязуетесь исправить найденные нами ошибки, недочеты и реализовать какие-то небольшие пожелания, если они будут.

Все исключительные и иные права на данную разработку и все материалы по проекту принадлежат заказчику в полном объеме и без ограничений.