Федеральное государственное образовательное учреждение высшего образования

«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

**Департамент анализа данных, принятия решений и финансовых технологий**

**В.И. СОЛОВЬЕВ, А.А. КОЧКАРОВ**

ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

**Сборник заданий с методическими указаниями для самостоятельной работы студентов**

Для студентов, обучающихся

по направлению подготовки 09.04.03 «Прикладная информатика», Магистерская программа «Интеллектуальные информационные технологии в управлении финансами организации»

**Москва – 2017**

Федеральное государственное образовательное учреждение высшего образования

«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

**Департамент анализа данных, принятия решений и финансовых технологий**

**В.И. СОЛОВЬЕВ, А.А. КОЧКАРОВ**

Технологии обработки больших данных

**Сборник заданий с методическими указаниями для самостоятельной работы студентов**

Для студентов, обучающихся

по направлению подготовки 09.04.03 «Прикладная информатика», Магистерская программа «Интеллектуальные информационные технологии в управлении финансами организации»

Одобрено на заседании Совета департамента

анализа данных, принятия решений и финансовых технологий (протокол № от 2017 г.)

Москва – 2017

**УДК 681.3**

**ББК 32**

**С 60**

**Рецензент:** Макрушин С.В., к.э.н., доцент Департамента анализа данных, принятия решений и финансовых технологий.

**Соловьев В.И., Кочкаров А.А. «Технологии обработки больших данных»** / Сборник заданий с методическими указаниями для самостоятельной работы студентов. – М.: Финансовый университет при Правительстве РФ, Департамент анализа данных, принятия решений и финансовых технологий, 2017. – 35 с.

Сборник заданий с методическими указаниями для самостоятельной работы студентов для студентов, обучающихся по направлению подготовки 09.04.03

«Прикладная информатика», Магистерская программа «Интеллектуальные информационные технологии в управлении финансами организации».

*УДК* ***681.3***

*ББК 32*

С 60

*Учебное издание*

**Соловьев Владимир Игоревич Кочкаров Азрет Ахматович**

**«Технологии обработки больших данных»**

Сборник заданий для самостоятельной работы студентов

Компьютерный набор **Кочкаров А.А.** Компьютерная верстка **Кочкаров А.А.** Формат 60х90/16. Гарнитура Тimes New Roman Усл. п.л. 4,0.

**© Финуниверситет, 2017**

**© В.И. Соловьев, 2017**

**© А.А. Кочкаров, 2017**

***Оглавление***

[Предисловие. Вводный обзор 5](#_bookmark0)

[Цели 9](#_bookmark1)

[Упражнения 10](#_bookmark2)

[Упражнение 1. Создание концентратора событий для ввода 10](#_bookmark3)

[Упражнение 2. Отправка событий в центр событий 15](#_bookmark4)

[Упражнение 4: Подготовка запросов и тестирование с помощью выборочных](#_bookmark5) [данных 30](#_bookmark5)

[Упражнение 5: Создание концентратора событий для вывода 39](#_bookmark6)

[Упражнение 6: Создание панели управления в реальном времени 48](#_bookmark7)

[Упражнение 7: Анализ потока данных в реальном времени 59](#_bookmark8)

[Заключение 63](#_bookmark9)

# Предисловие. Вводный обзор

Машинное обучение (МО) — подобласть компьютерной науки, которая занимается такими задачами, как распознавание лиц, распознавание речи, обработка естественного языка, прогнозный анализ, глубокое обучение (глубокие нейронные сети) и многие другие. Для решения таких задач применяются различные инструменты. Так как существует много инструментов машинного обучения стоит их разделить на платформы и библиотеки. Платформа предоставляет все, что нужно для запуска проекта, в то время как библиотека предоставляет только дискретные возможности. В нынешний день существуют различные инструменты и платформы для решения задач машинного обучения. Онлайн-платформы могут предоставить все те возможности, которые доступны в инструментах машинного обучения и даже больше, так как разработчиками таких платформ являются лучшие специалисты в этой сфере. Такие платформы предоставляют возможности для завершения проекта машинного обучения от начала до конца. А именно, анализ данных, подготовка данных, моделирование, оценка и выбор алгоритма.

1. ***IBM Watson***

Платформа компании IBM предлагает возможности прогнозного анализа и визуализации данных, и диалоговый тип интерфейса. Она автоматически делает сложную математику, чтобы показать наиболее значимые факты, закономерности и отношения. Бесплатная версия предлагается с ограничениями на объемы данных. Сейчас на основе этой облачной API-платформы разрабатываются различные приложения, сервисы и процессы. У Watson на данный момент 28 API, и в основе всего этого лежит более 50 различных технологий. К концу 2016 года разработчики обещают добавить еще пару десятков API, и к 2017 году их число должно увеличиться до 50.

Примеры применения:

**Машинный перевод -** помогает переводить текст на разные языковые пары. **Message Resonance -** определения популярности фразы или слова с заранее определенной аудиторией.

**«Вопросы и ответы» -** Эта служба обеспечивает прямые ответы на запросы, инициируемые первичными источниками документов.

**User Modelling –** предоставляет возможность делать прогнозы относительно характеристик кого-либо из данного текста.

1. ***Microsoft Azure Machine Learning API***

Azure Machine Learning API - инфраструктура облачных вычислений от компании Microsoft. В контексте

машинного обучения инструмент представляет собой платформу, которая позволяет вне зависимости от качества данных строить решения прямо «на облаке» и с легкостью создавать на их основе BI-приложения

Azure Machine Learning API предоставляет различные функции, такие как:

Возможность создания настраиваемого конфигурируемого R-модуля, чтобы аналитики данных или ученые, работающие с данными, могли использовать свой собственный код языка R для обучения или прогнозирования задач.

Azure позволяет ученым или аналитикам данных включать свои собственные скрипты Python с использованием различных библиотек Python для науковедения, таких как SciPy, SciKit-Learn, NumPy, Pandas и т. Д. Также могут использоваться популярные инструменты, такие как iPython Notebook и другие инструменты Python для Visual Studio С Azure Machine Learning API.

Azure Machine Learning API также поддерживает Spark и Hadoop для большой обработки данных, что делает его идеальным выбором независимо от платформы.

1. ***Google Prediction API***

Данный сервис облачных вычислений от Google предназначен для работы с большими массивами данных и машинного обучения на основе результатов их анализа, с его помощью можно прогнозировать изменяющиеся процессы в будущем, например, посещаемость вашего сайта.

Prediction API предоставляет возможности машинного обучения и, учитывая наборы данных, на основе готовых шаблонов можно создавать приложения, которые будут выполнять, например, такие задачи:

учитывая прошлые привычки просмотра пользователя, предсказывать, какие фильмы или продукты пользователю могут понравиться в будущем; классифицировать электронную почту как спам или не спам;

анализировать комментарии о продукте или компании и определять, являются ли они положительными или отрицательными;

предугадывать, сколько пользователь может потратить в конкретный день, учитывая его историю расходов.

Google Prediction API [6] может интегрироваться с App Engine, и RESTful. API доступна через библиотеки для многих популярных языков, таких как Python, JavaScript и.NET.

1. ***Amazon Machine Learning API***

это сервис, который предоставляет возможности для разработчиков всех уровней квалификации используя технологию машинного обучения. Amazon Machine Learning предоставляет инструменты визуализации и мастеров, которые ведут вас через процесс создания модели машинного обучения без необходимости изучения сложных алгоритмов и технологий МО. После того, как модели будут готовы, Amazon Machine Learning позволяет легко получить предсказания для приложения, используя простые API-интерфейсы, без необходимости реализации пользовательского кода.

**Примеры использования API Amazon Machine Learning API**

Используя API AWS, ученые или аналитики могут классифицировать жанр песни, анализируя характеристики уровня сигнала.

Он может использоваться для распознавания человеческой деятельности на основе акселерометра, данных датчиков, записанных на смартфонах или сигналах гироскопа, чтобы узнать, идет ли мобильное устройство вверх по лестнице, идет ли вниз по вертикали или по горизонтали, сидя или стоя.

Анализ записей активности веб-сайтов для обнаружения поддельных пользователей, ботов или спамеров в системе.

1. ***BigML***

это платформа машинного обучения на основе облака с простым в использовании графическим интерфейсом. Она также предоставляет простые механизмы для включения прогнозирующих моделей в приложениях через REST API. Платформа включает в себя наблюдаемое обучение (для построения прогнозных моделей), обучения без учителя (для понимания поведения), обнаружения аномалий (используется в выявлении случаев мошенничества), средства визуализации данных (разброс-диаграмм и графиков Sunburst) и множество механизмов для изучения данных. Скромное ценообразование делает её привлекательным для среднего и крупного бизнеса, которые хотят преимущества в работе с машинным обучением без больших предварительных затрат и задержек в реализации. BigML является прагматичной, с низкой стоимостью, простой в использовании платформой для построения мощных прогностических моделей.

*Примеры использования BigML:*

BigML помогает создать описательную модель, чтобы понять отношения между различными атрибутами сложных данных и прогнозируемыми атрибутами, чтобы бизнес-аналитики могли играть в сценарии «что-если».

BigML полезен, если аналитик хочет создать прогностическую модель, основанную на прошлых примерах. Аналитики могут построить прогнозирующую модель, используя BigML, чтобы сделать предсказание в пакетном режиме, которое можно загрузить как файл CSV для анализа.

BigML также находит применение, когда необходимо делать периодические прогнозы. Модель может поддерживаться на платформе BigML, а затем может использоваться удаленно по мере необходимости с использованием интерфейса командной строки BigMLer.

1. ***Splunk***

Splunk использует методы машинного обучения для выявления аномалий и повторяемых элементов, которые могут ускорить расследование инцидентов и получение аналитической информации. Этот встроенный специализированный тип машинного обучения помогает обнаруживать тенденции и всплески, а также устранить «шум», создаваемый огромным количеством событий в данных. Splunk предоставляет все возможности машинного обучения посредством расширенной аналитики в своих платных решениях.

**Расширенные средства аналитики Splunk IT Service Intelligence включают:**

**Обнаружение аномалий:** определение базовых повторяемых элементов нормальной работы и статистические измерения для выявления непостоянных повторяемых элементов пороговых значений.

**Адаптация пороговых значений:** динамическая подстройка пороговых значений к изменениям поведения и выделение аномальной активности.

**Анализ событий:** комбинация данных событий с расширенной аналитикой для уменьшения загромождения журналов, числа ложных срабатываний и поддержки масштабных правил.

***Обработка больших данных с помощью Azure Stream Analytics***

Azure Stream Analytics – это облачная служба для приема высокоскоростных потоков данных с устройств, датчиков, приложений, веб- сайтов и других источников данных и анализа этих данных в режиме реального времени. Он поддерживает SQL-подобный язык запросов, который работает над динамическими потоками данных и делает анализ постоянно изменяющихся данных не сложнее, чем выполнение запросов к статическим данным, хранящимся в традиционных базах данных. С помощью Azure Stream Analytics вы можете настроить задания, анализирующие входящие данные, на аномалии или информацию, представляющие интерес, и записывать результаты, представлять уведомления на информационных панелях или даже отправлять предупреждения на мобильные устройства. И все это можно сделать при низкой стоимости и с минимальными усилиями.

Сценарии применения аналитики данных в реальном времени являются легитимными и включают в себя обнаружение мошенничества, защиту от кражи идентификационной информации, оптимизацию распределения ресурсов (например транспортная служба типа Uber, которая отправляет водителей в районы, где спрос возрастает до того, как спрос достиг пика), Анализ потока кликов на веб-сайтах, предложения по покупкам на сайтах розничной торговли и множество других. Имея возможность обрабатывать данные по мере их поступления, а не дожидаться, пока они будут объединены, дает преимущества для компаний, которые способны вносить корректировки «на лету».

В этой лабораторной работе вы создадите задание Azure Stream Analytics и используйте его для анализа данных с имитируемых устройств Internet Things (IoT). Вы увидите, как просто контролировать потоки данных в реальном времени для получения важной информации. И в конце вы построите веб- приложение, которое будет отображать результаты задания Stream Analytics в реальном времени.

# Цели

В процессе этой самостоятельной работы должно бытьи зучено как:

* Создать концентраторы событий Azure и использовать их для входа и выхода Stream Analytics.
* Создать задание Stream Analytics и проверить запросы на потоки выборочных данных.
* Запустить задание Stream Analytics и выполнить запросы к потокам данных в реальном времени.
* Использовать события из концентратора событий и отображать их в веб- приложении.

Для выполнения всех упражнений потребуется подписка или бесплатная регистрация на Microsoft Azure.

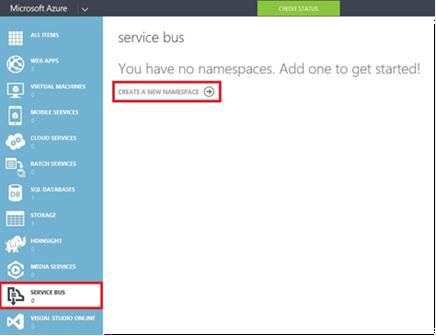
# Упражнения

## Упражнение 1. Создание концентратора событий для ввода

Azure Stream Analytics поддерживает несколько типов ввода, в том числе входные данные от хранилищ BLOB-объектов Azure и входные данные от концентраторов событий Azure. Использование концентраторов событий Azure как правило, более интересно, потому что в мире IoT данные легко передаются концентраторам событий Azure через полевые шлюзы (для устройств, которые не поддерживают IP) или облачные шлюзы (для устройств, поддерживающих IP) И один концентратор событий Azure может обрабатывать миллионы событий в секунду, передаваемых с устройств, разбросанных по всему миру

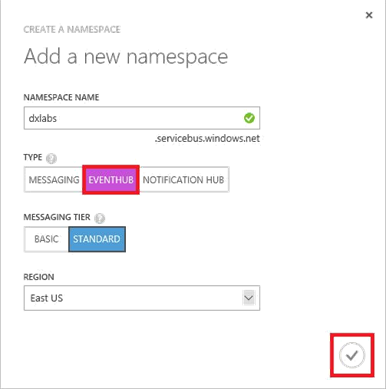
В этом упражнении вы создадите концентратор событий Azure для ввода данных в Azure Stream Analytics и настройки его таким образом, чтобы к ним можно было безопасно и безопасно обращаться с помощью устройств IoT и шлюзов

Поскольку в настоящее время Azure Portal не имеет полной поддержки концентраторов событий и Stream Analytics, вы будете работать в этой лаборатории с помощью классического портала. Перейдите на страницу https://manage.windowsazure.com, чтобы открыть классический портал, и нажмите «Service Bus» на ленте слева. Затем нажмите CREATE A NEW NAMESPACE (см.рис 1.1) для создания нового пространства имен служебной шины. (Если вы уже создали одно или несколько пространств имен, нажмите + NEW в нижнем левом углу страницы, чтобы создать еще одно.)



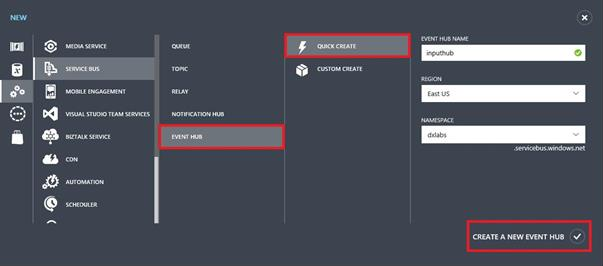
**Рис.1.1. Azure Service Bus**

Введите имя пространства имен в поле NAMESPACE NAME. Имя должно быть уникальным в пределах Azure (Зеленая галочка появится в поле, когда введенное вами имя будет доступным к созданию.) Установите TYPE в EVENTHUB и выберите ближайший к вам регион из раскрывающегося списка REGION. Затем щелкните галочку в правом нижнем углу диалогового окна (см. рис. 1.2).

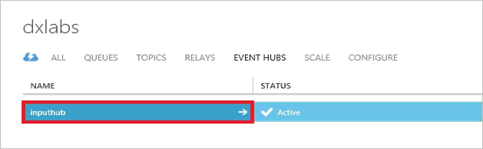


**Рис.1.2. Создание пространства имен служебной шины**

Нажмите кнопку + NEW в левом нижнем углу страницы. Нажмите EVENT HUB, затем QUICK CREATE. Введите «inputhub» (без кавычек) в поле EVENT HUB NAME (имя не обязательно должно быть уникальным внутри Azure). Выберите тот же самый регион, который вы выбрали для пространства имен служебной шины на предыдущем шаге, и убедитесь, что пространство имен, созданное на этом шаге, выбрано в поле NAMESPACE. Затем нажмите CREATE A NEW EVENT HUB в нижнем правом углу (см. 1.3).



**Рис.1.3. Создание центра событий**

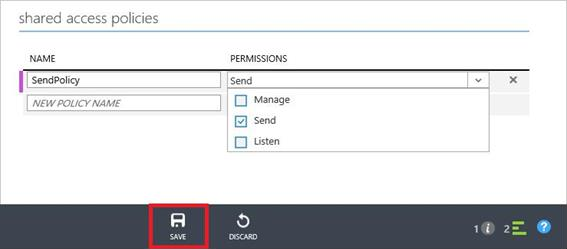


**Рис.1.4. Открытие центра событий**

Дождитесь создания концентратора событий. Затем щелкните имя концентратора событий, чтобы отобразить панель мониторинга событий. Нажмите CONFIGURE (см. рис. 1.5).

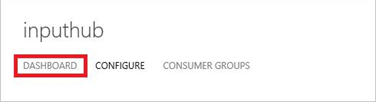
**Рис.1.5. Настройка концентратора событий**

Чтобы передать события концентратору событий из приложения или устройства, вам необходимо создать политику общего доступа, которая включает разрешение отправки. В разделе политики общего доступа на странице создайте новую политику, введя «SendPolicy» (без кавычек) в первое текстовое поле поставив галочку «Send» в раскрывающемся списке

«Permissions». Затем нажмите кнопку «SAVE» внизу страницы, чтобы сохранить новую политику (см. рис. 1.6).

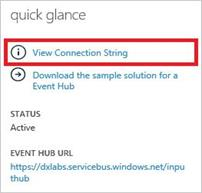
**Рис.1.6. Создание новой политики**

Нажмите DASHBOARD в верхней части страницы, чтобы вернуться к панели инструментов.



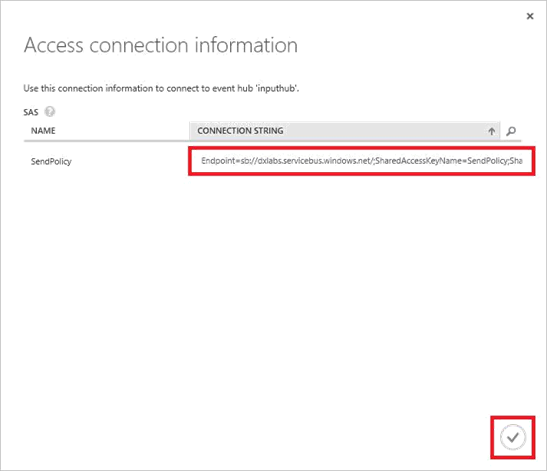
**Рис.1.7. Возвращение к панели инструментов**

Под quick glance в правой части страницы, нажмите View Connection String (см. рис. 1.8.).



**Рис.1.8. Просмотр строки подключения**

Скопируйте строку соединения в буфер обмена. (Если вы наведите курсор на строку соединения, появится кнопка копирования в буфер обмена) Затем щелкните галочку, чтобы закрыть диалог (см. рис.1.9). Завершите работу, вставив строку соединения в ваш любимый текстовый редактор, чтобы вы могли найти ее в следующем упражнении.



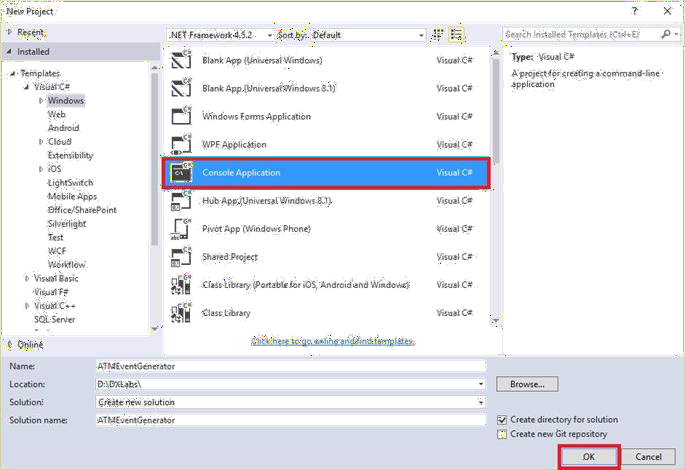
**Рис.1.9. Копирование строки подключения**

Вы создали концентратор событий, который может принимать события и использоваться в качестве источника ввода в задание Stream Analytics. Кроме того, вы создали политику, позволяющую владельцам этой политики отправлять события на концентратор событий и строку подключения, в которой эта политика инкапсулируется. Следующий шаг – использовать эту строку соединения для передачи событий в концентратор событий.

## Упражнение 2. Отправка событий в центр событий

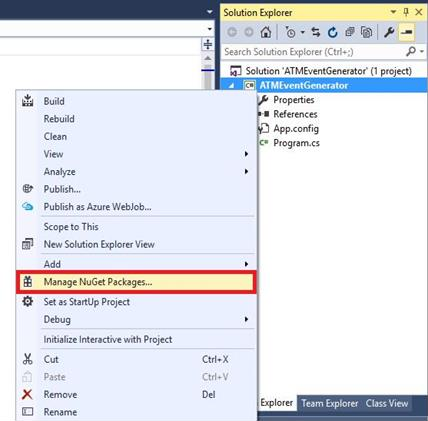
Приложения, устройства и шлюзы могут отправлять события в концентраторы событий Azure с помощью Azure REST API концентраторов событий Azure или расширенного протокола очереди сообщений (коротко AMQP). В этом упражнении вы напишете консольное приложение Windows, которое использует AMQP для отправки событий на концентратор событий, созданный вами в упражнении 1. Каждое событие будет представлять собой изъятие из банкомата и будет содержать соответствующую информацию, такую как номер карты, снятие, время и сумма снятия и уникальный идентификатор используемого банкомата.

Запустите Visual Studio 2015 и нажмите File → New → Project чтобы создать новое приложение консоли Windows с именем «ATMEventGenerator» (см. рис. 2.1).



**Рис.2.1 Создание нового консольного приложения**

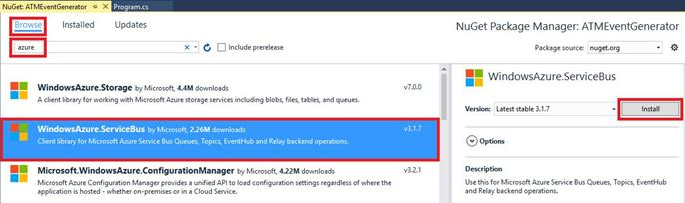
В окне «Solution explorer» щелкните правой кнопкой мыши проект ATMEventGenerator и выберите «Manage NuGet Packages» (см. рис. 2.2).



**Рис.2.2 Управление пакетами NuGet для проекта**

NuGet – это бесплатный менеджер пакетов с открытым исходным кодом для платформ разработки Microsoft. Он предоставляет доступ к тысячам библиотек или пакетов, содержащих код для выполнения множества задач. Он интегрирован в Visual Studio 2015, что позволяет легко добавлять пакеты NuGet в ваш проект и совершать много чего, не написав много кода.

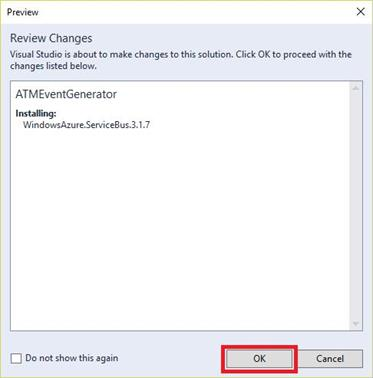
Нажмите Browse. Затем введите «azure» (без кавычек) в поле поиска. Нажмите WindowsAzure.ServiceBus для выбора пакета сервисной шины Azure от NuGet. Наконец, нажмите «Install», чтобы установить последнюю стабильную версию пакета (см. рис. 2.3).. Этот пакет содержит API, которые ваше приложение будет использовать для отправки событий в концентратор



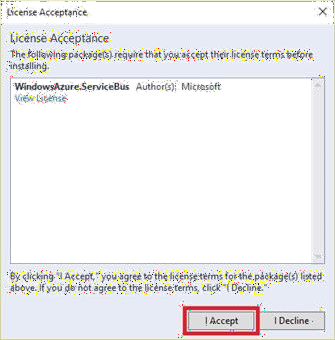
событий.

**Рис.2.3 Установка WindowsAzure. ServiceBus**

Если будет предложено просмотреть изменения, нажмите кнопку «ОК». Опционально установите флажок «Do not show this again», чтобы не выводить запрос снова (см. рис. 2.4).

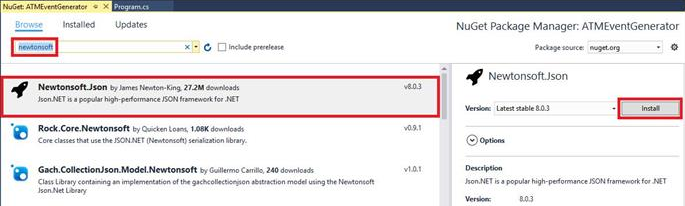


**Рис.2.4 Обзор изменений**

Если будет предложено принять лицензию для пакета нажмите «I Accept» (см. 2.5).

**Рис.2.5 Принятие лицензии на пакет**

Вернитесь в окно поиска и введите «newtonsoft», снова без кавычек. Выберите Newtonsoft.Json и нажмите Install, чтобы установить последнюю

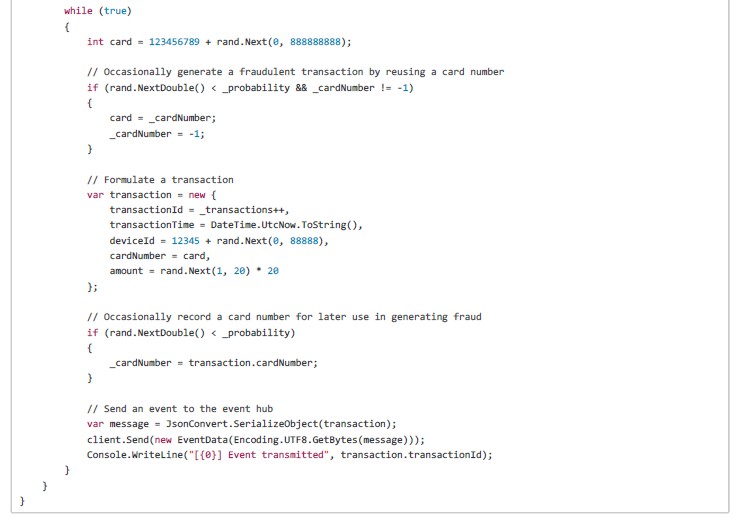
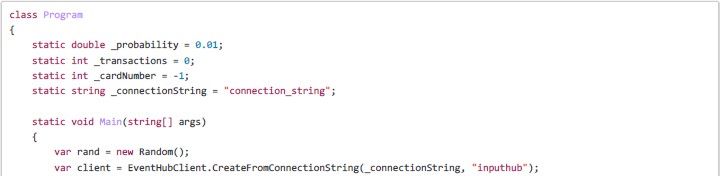
стабильную версию Json.NET. Этот пакет содержит удобные API для генерации и использования JSON. Если будет предложено просмотреть изменения, нажмите OK.

**Рис.2.6 Установка Json.NET**

Откройте Program.cs и добавьте следующие операторы в using statements, расположенные в верхней части файла:



Замените пустой класс программы следующей реализацией:



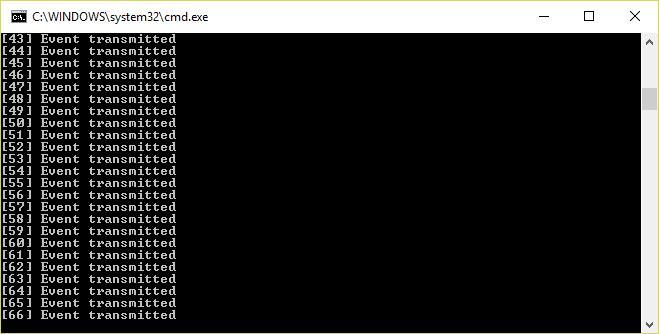
Замените connection\_string в строке 6 с помощью строки подключения, которую вы скопировали в буфер обмена (и, надеюсь, в ваш любимый текстовый редактор) в упражнении 1, шаг 9.

Удалите ";EntityPath=inputhub" от конца строки подключения. The



\_connectionString поле должно выглядеть примерно так:

Нажмите Ctrl+F5 для запуска программы. Убедитесь, что вы получили вывод, подобный следующему. Каждая строка представляет одно событие, отправленное на концентратор событий, и события, вероятно, будут скатываться со скоростью около 2 - 3 в секунду. (Ставки будут варьироваться в зависимости от скорости вашего соединения.) Убедитесь, что никаких исключений не выбрано.

**Рис.2.6 Сообщения, отправленные в центр событий**

Нажмите Ctrl+C чтобы остановить поток событий. Затем нажмите

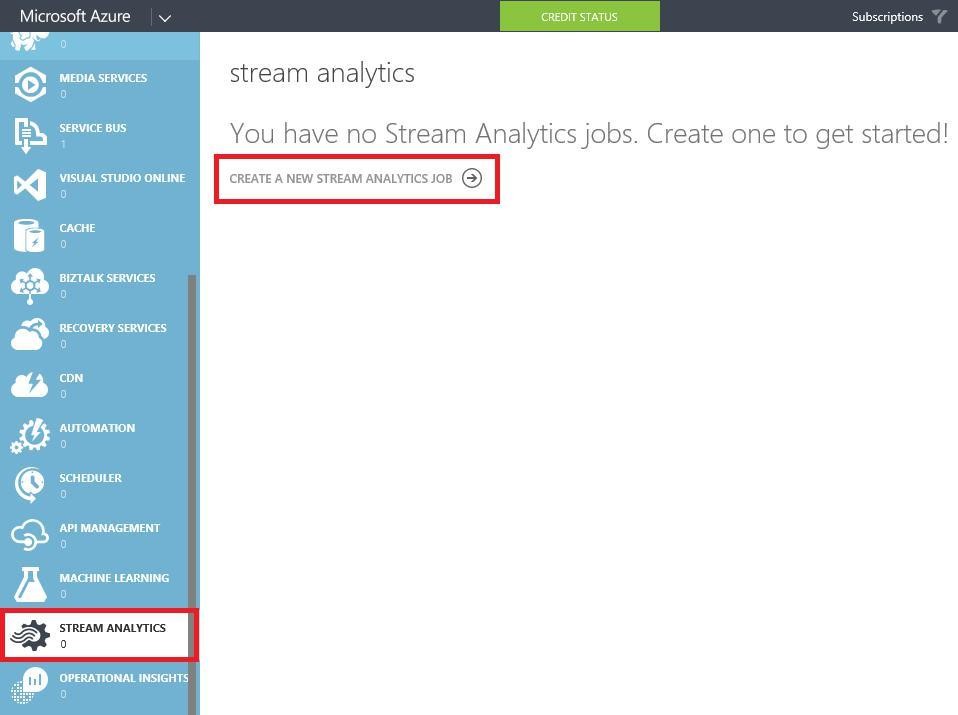
«Enter», чтобы закрыть окно «Командная строка».

В реальной жизни не было бы реальных банкоматов, отправляющих события в центр событий. ATMEventGenerator имитирует эти события в программном обеспечении и создает основу для создания и тестирования задания Stream Analytics, которое анализирует эти события.

***Упражнение 3: Создание задания Stream Analytics***

Теперь у вас есть программное обеспечение, которое отправляет события на концентратор событий Azure, и концентратор событий, который поглощает данные. В этом упражнении вы будете использовать портал Microsoft Azure Classic Portal для создания задания Stream Analytics и подключения его к концентратору событий. Вы также захватите исходные данные, представляемые Stream Analytics, с помощью концентратора событий и изучите его структуру.

Откройте Classic Portal в своем браузере, если он еще не открыт. Нажмите

«STREAM ANALYTICS» на ленте слева, а затем нажмите «CREATE A NEW STREAM ANALYTICS JOB».

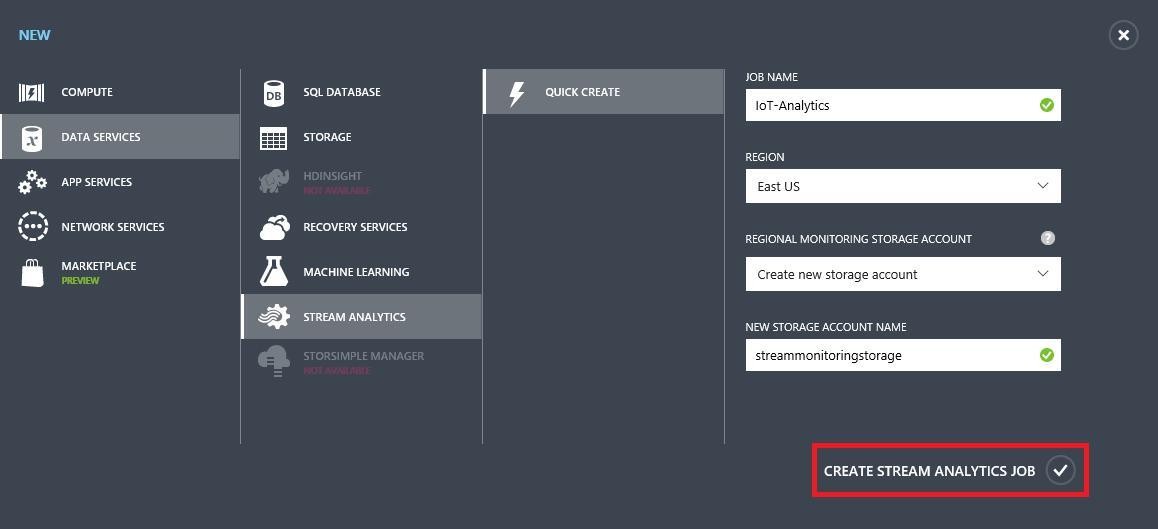
**Рис 3.1 Azure Stream Analytics**

Введите «IoT-Analytics» (без кавычек) в поле JOB NAME. Выберите ближайший к вам регион в поле REGION. (Здесь важно выбрать тот же регион, который вы выбрали для концентратора событий в Упражнении 1, потому что вы не платите за данные, перемещаемые в центре обработки данных, но вы платите за данные, перемещаемые между центрами данных. Кроме того , Поиск служб, которые говорят друг с другом в одном центре обработки данных, сокращает время ожидания). В разделе REGIONAL MONITORING STORAGE ACCOUNT укажите имя новой учетной записи хранилища или выберите

существующую учетную запись хранилища, если портал предоставит вам эту опцию

Если вы решите создать новую учетную запись хранилища, напомните, что имена учетных записей хранилища могут иметь длину от 3 до 24 символов, могут содержать только цифры и строчные буквы и должны быть уникальными в пределах Azure. Зеленая галочка рядом с названием указывает, что она соответствует всем этим критериям. Кроме того, рекомендуется размещать учетную запись хранилища в том же регионе, что и задание Stream Analytics, чтобы предотвратить перемещение данных между центрами обработки данных..

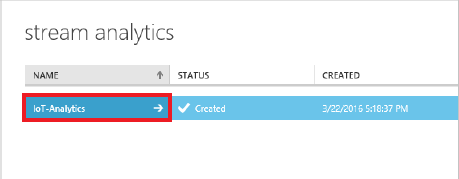
Когда все будет готово, нажмите CREATE STREAM ANALYTICS JOB в нижнем правом углу.



**Рис 3.2 Создание задания Stream Analytics**

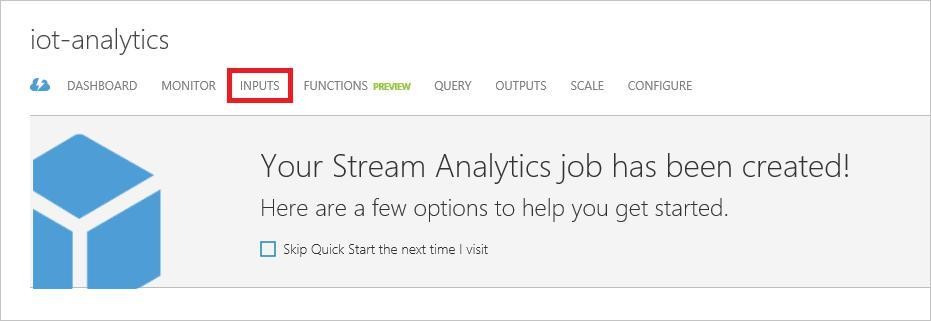
Через несколько секунд в портале появится задание Stream Analytics.

Подождите, пока задание не будет создано, а затем щелкните его.

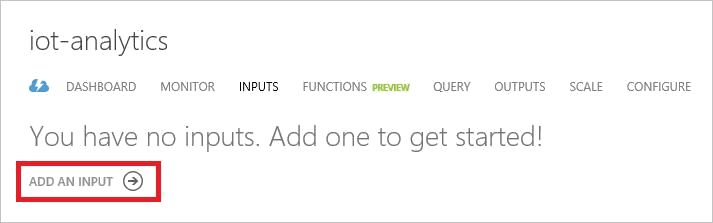


**Рис 3.3 Новое задание Stream Analytics**

Нажмите INPUTS в верхней части страницы.



**Рис 3.4 Странице IoT-аналитики**

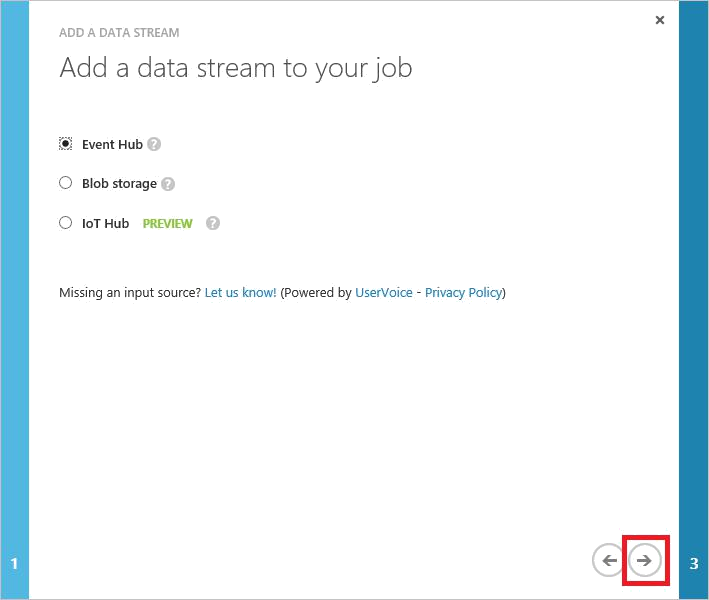
Нажмите ADD AN INPUT.

**Рис 3.5 Добавление ввода**

Убедитесь, что Data stream выбрано и щелкните правой стрелкой в нижнем правом углу диалогового окна.

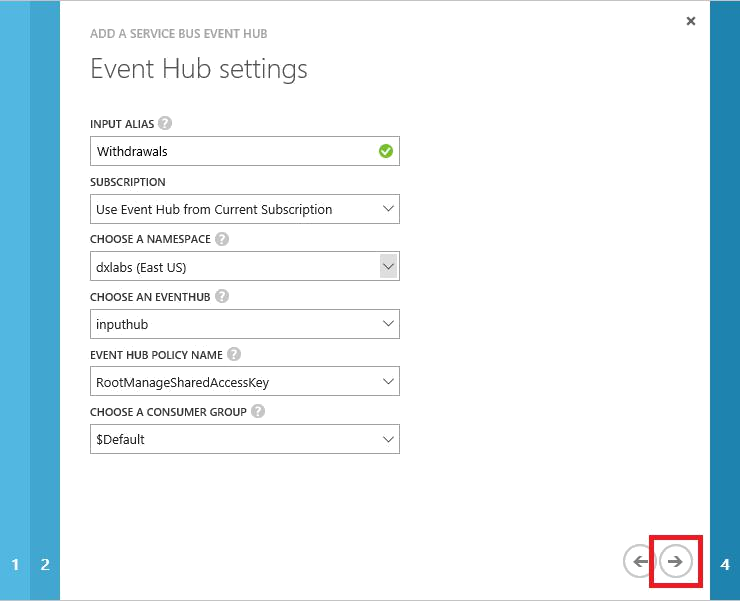


**Рис 3.6 Указание типа ввода**

Убедитесь, что Event Hub выбрано и затем нажмите на стрелку вправо.

**Рис 3.7 Указание типа потока данных**

Введите «Withdrawals» (без кавычек) в качестве дружественного псевдонима для ввода в поле INPUT ALIAS. В блоках CHOOSE A NAMESPACE и CHOOSE AN EVENTHUB выберите пространство имен и концентратор событий, которые были созданы в упражнении 1. Установите параметр политики безопасности для событий EVENT HUB POLICY NAME равным RootManageSharedAccessKey (это политика по умолчанию, которая автоматически создается при создании концентратора событий, предоставляет разрешения Для управления концентратором событий, отправки событий и получения событий) и and CHOOSE A CONSUMER GROUP, чтобы установить значение $Default. Затем нажмите правую стрелку в нижнем правом углу.



**Рис 3.8 Задание настроек event-hub**

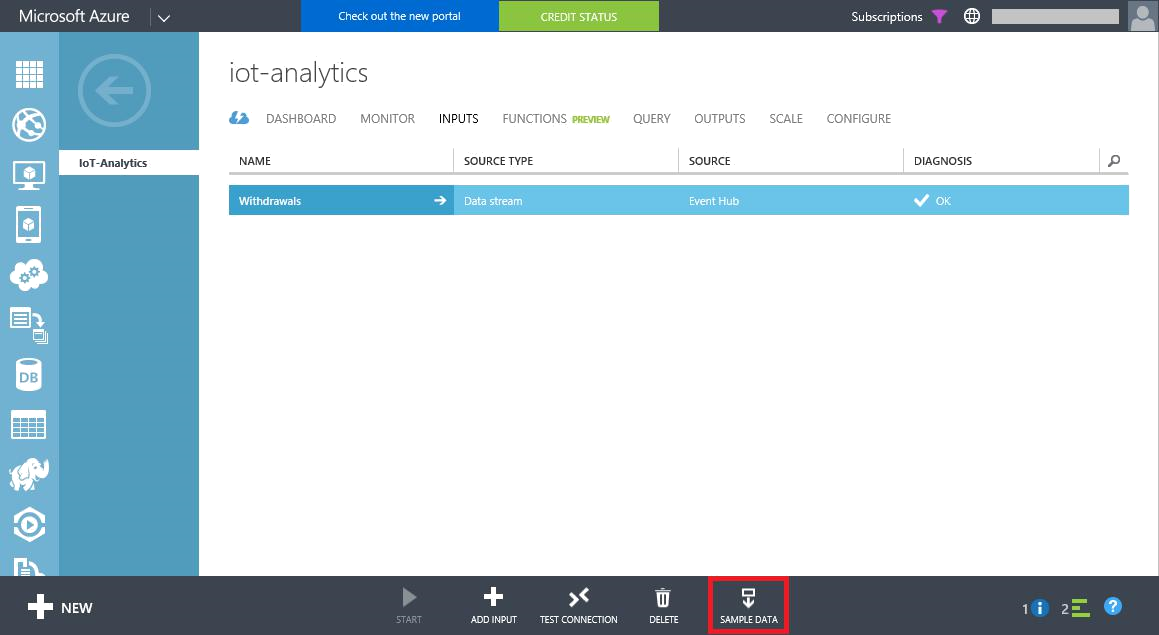
Убедитесь, что JSON выбран в EVENT SERIALIZATION FORMAT (консольное приложение, которое отправляет события в концентратор событий (event-hub), отправляет данные JSON), а UTF8 выбирается под ENCODING. Затем нажмите галочку в правом нижнем углу, чтобы завершить добавление ввода.

**Рис 3.9** Указание формата сериализации

Через несколько секунд новый ввод - «Withdrawals» - появится в списке входов для задания Stream Analytics. Вернитесь в Visual Studio и снова запустите ATMEventGenerator..

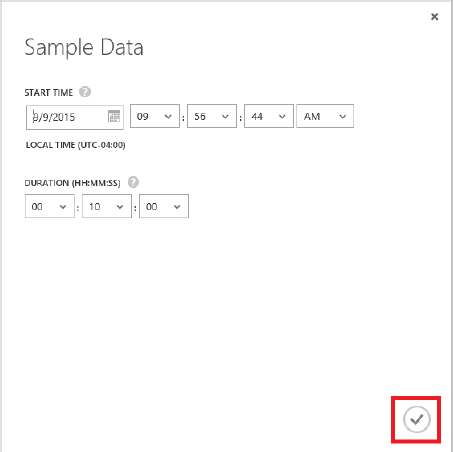
Разрешите ATMEventGenerator поработать несколько минут. Затем нажмите Ctrl+C для его остановки и Enter, чтобы закрыть его, и вернуться в портал, открытый в вашем браузере.

Нажмите кнопку SAMPLE DATA в нижней части страницы для выборки данных из центра событий.



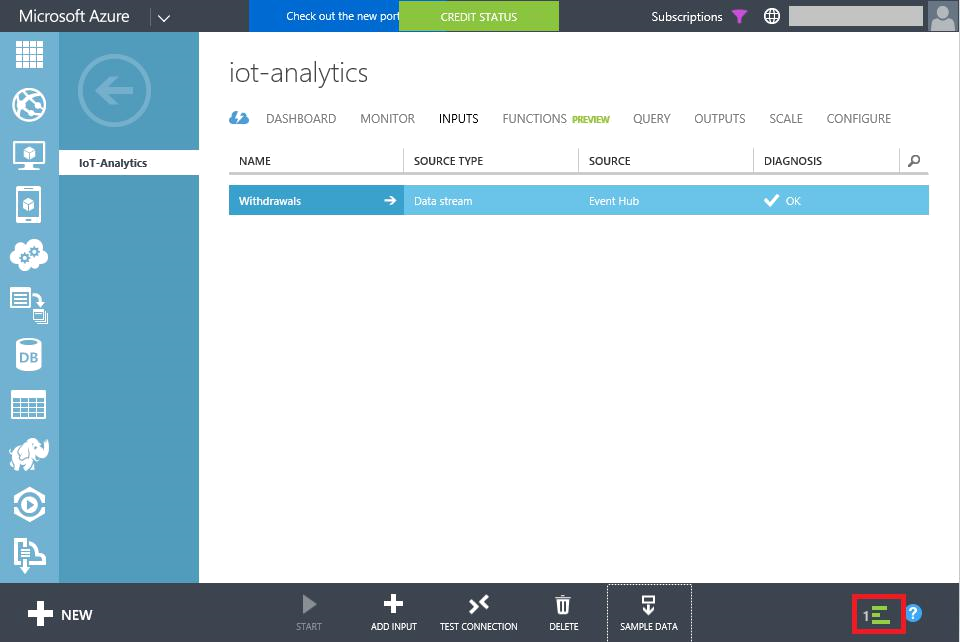
**Рис 3.10 Входные данные выборки**

Нажмите галочку в нижнем правом углу следующего диалога, чтобы отобрать любые данные, переданные на концентратор событий за последние 10 минут. (Вот почему вы снова запустили ATMEventGenerator: чтобы убедиться в наличии данных для выборки, даже если прошло более 10 минут после завершения упражнения 2.).

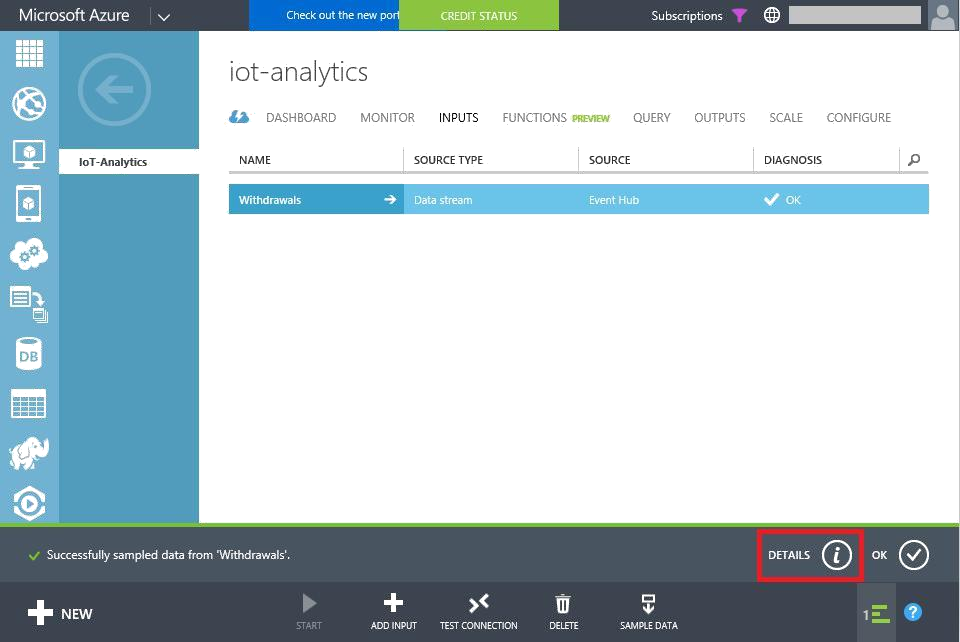


**Рис 3.11 Указание времени начала и продолжительности**

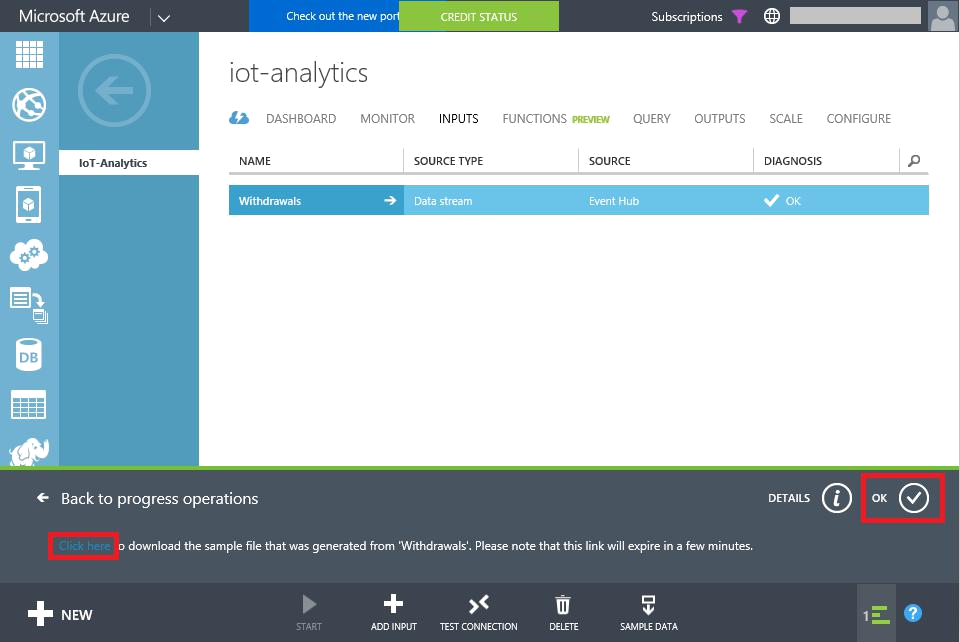
Подождите, пока выборка не будет завершена. Затем нажмите кнопку в правом нижнем углу страницы, которая указывает на завершение операции.



**Рис 3.12 Отбор данных завершен**

Когда появится лента с надписью «Successfully sampled data from Withdrawals», нажмите кнопку «Details» справа.

**Рис 3.13 Выборка данных выполнена успешно**

Нажмите Click here чтобы загрузить данные, взятые из концентратора событий. Сохраните файл JSON, загруженный в место, где его можно легко найти. Затем нажмите «ОК», чтобы закрыть ленту.

**Рис 3.14 Загрузка образцов данных**

Откройте файл JSON, загруженный в ваш любимый текстовый редактор, и найдите момент, чтобы изучить его содержимое. Сколько строк (событий) представлено в выборке данных? Какова структура каждой строки - то есть, какие поля содержит каждая строка?

Если вы обнаружите, что вывод трудно переварить, поскольку нет разрывов строк, попробуйте вставить его в онлайн-просмотрщик JSON, например, на <http://jsonviewer.stack.hu/> или https://jsonformatter.curiousconcept.com/..

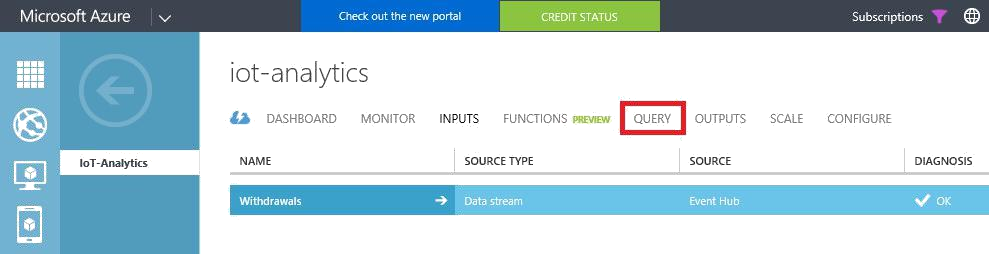
Итак, вы подключили задание Stream Analytics к концентратору событий и продемонстрировали, что данные передаются от одного к другому. Вы также изучили структуру этих данных. Следующим шагом будет сделать что-то с этим, в частности, использовать Azure Stream Analytics для переноса данных.

## Упражнение 4: Подготовка запросов и тестирование с помощью выборочных данных

Теперь, когда ваша работа настроена, вы можете сделать с Stream Analytics намного больше, чем просто просматривать исходные данные, представленные ей. Весь смысл Stream Analytics заключается в возможности выполнения запросов к данным, даже несмотря на то, что данные динамические, а не статические. В этом упражнении вы будете использовать язык запросов аналитики Stream Analytics для запроса образца набора данных для потенциально мошеннических транзакций в банкоматах. Рекомендуется всегда тестировать свои запросы на выборочных данных перед их развертыванием в потоках данных в реальном времени, так как с образцовыми данными вы можете проверить, что известный набор входов дает ожидаемый набор выходов.

Чтобы пометить потенциально мошеннические изъятия из банкоматов, вы будете запрашивать транзакции, выполненные с одной и той же карточкой ATM, в разных банкоматах в течение заданного временного окна (60 секунд). В реальной жизни вы, вероятно, использовали бы большее окно времени и, возможно, даже фактор расстояния между банкоматами. Тем не менее, более узкое временное окно полезно в лабораторной среде, поскольку оно позволяет выполнять значимые эксперименты в течение нескольких минут, а не часов.

Начните с возврата к заданию Stream Analytics на портале и нажмите



QUERY в верхней части страницы.

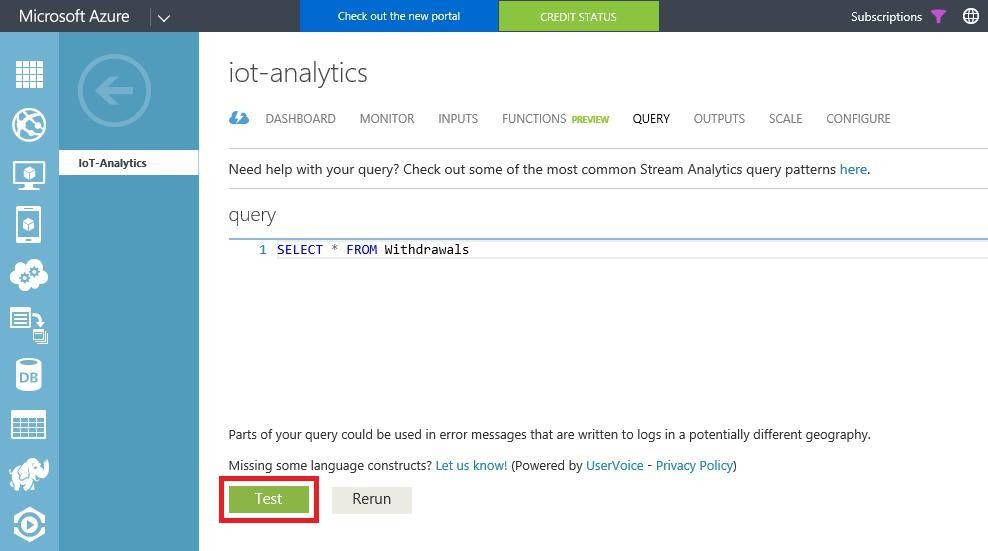
**Рис 4.1 Переход на страницу запроса**



Введите следующий запрос в поле запроса и нажмите кнопку «Test».

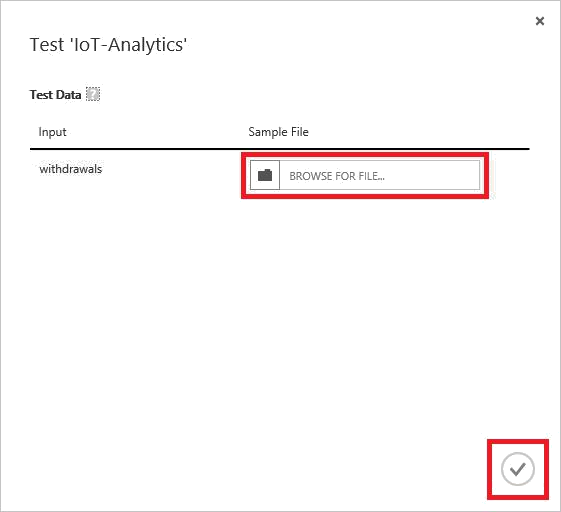
Откуда взялось название «Withdrawals»? Это псевдоним, который вы назначили для входа event-hub в предыдущем упражнении. Если вы назвали его

по-другому, вам нужно заменить «Withdrawals» на псевдоним, который вы использовали.



**Рис 4.2 Тестирование запроса**

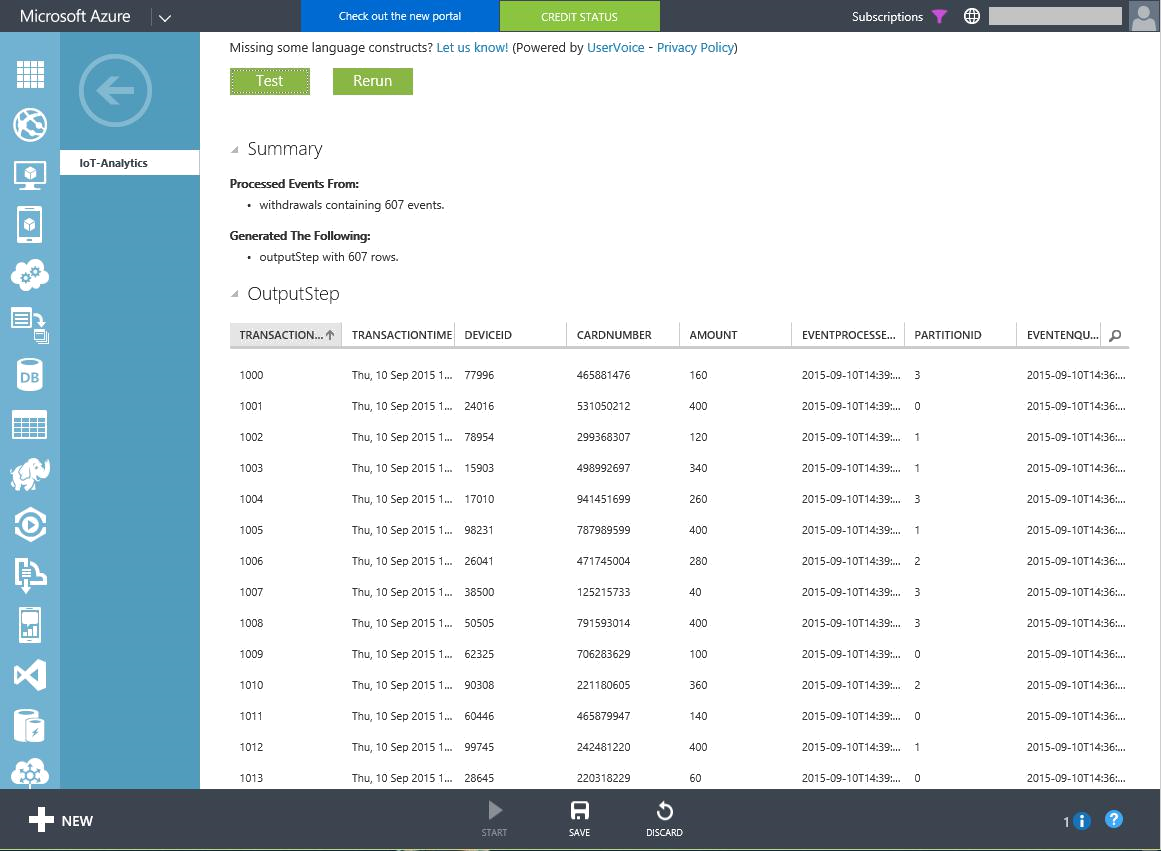
В появившемся диалоговом окне нажмите BROWSE FOR FILE. Выберите файл Withdrawals.json, указанный в каталоге «resources» этой лаборатории. Затем нажмите OK, щелкнув галочку в нижнем правом углу диалогового окна.



**Рис 4.3 Загрузка тестовых данных**

Прокрутите страницу вниз и убедитесь, что вы видите выход, изображенный ниже. Данные испытаний содержат 607 строк. В каждой строке есть поля с именами TRANSACTIONID, TRANSACTIONTIME, DEVICEID,

CARDNUMBER и AMOUNT. DEVICEID - это идентификатор банкомата, на котором была совершена транзакция. AMOUNT - сумма денег, изъятых из



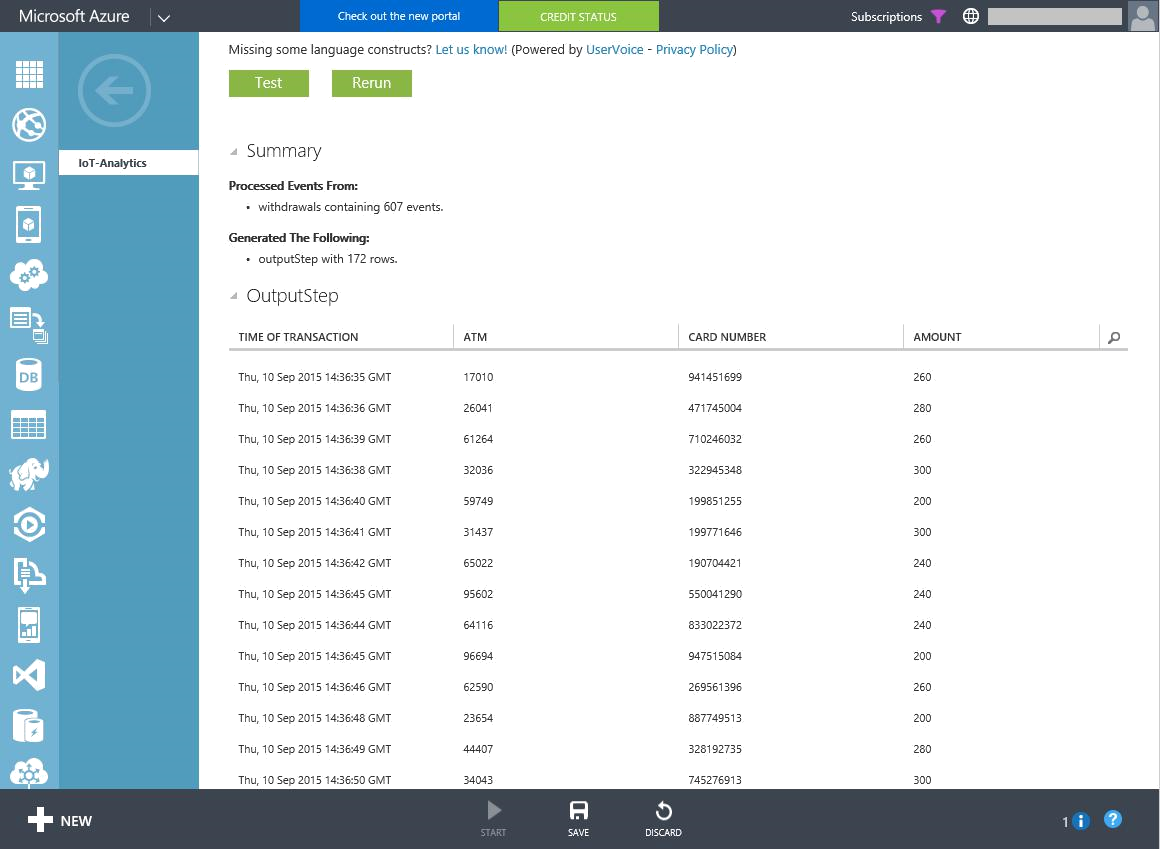
банкомата.

**Рис 4.4 Вывод SELECT**

Предположим, вы хотите просмотреть транзакции только для сумм от 200 до 300 включительно. Кроме того, предположим, что вы хотели очистить вывод, назначив ваши собственные имена столбцов и исключая столбец TRANSACTIONID. Введите следующий запрос и нажмите кнопку

«Повторить», чтобы проверить его. (Rerun выполняет запрос к уже загруженным тестовым данным. Если вы хотите загрузить другой тестовый файл, нажмите кнопку Test еще раз.).

Прокрутите вниз и проверьте, что запрос сформировал следующий



результат:

**Рис 4.5 Настройка вывода**

Одной из ключевых особенностей языка запросов Stream Analytics Query является возможность группировать результаты с использованием окон времени, длину которых вы указываете. Введите следующий запрос, чтобы подсчитать количество транзакций, происходящих каждую минуту, и нажмите

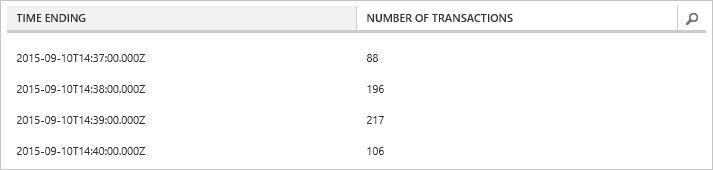


«Rerun» для его выполнения.

TIMESTAMP BY является важным элементом языка запросов Stream Analytics. Если в приведенном выше запросе не было указано, вы будете запрашивать количество транзакций, поступивших в концентратор событий каждую минуту, а не количество транзакций, которые произошли в каждом 1-

минутном интервале. TIMESTAMP BY позволяет указать поле во входном потоке как время события..

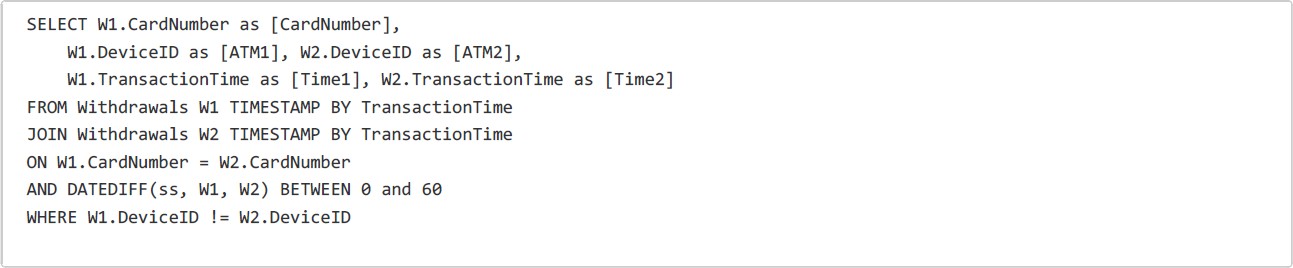
Прокрутите вниз и убедитесь, что вы видите нижеуказанный результат:



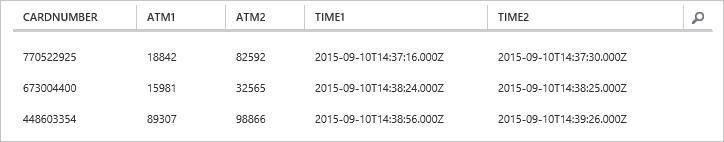
**Рис 4.6** Количество транзакций в минуту

Теперь пришло время запрашивать тестовые данные для потенциально мошеннических транзакций - транзакций с одной и той же карточкой ATM, но с разными банкоматами, которые происходят в течение 60 секунд друг от друга. Это запрос, который вы позже будете использовать для потока данных в реальном времени.

Введите следующий запрос и нажмите «Rerun», чтобы выполнить его:

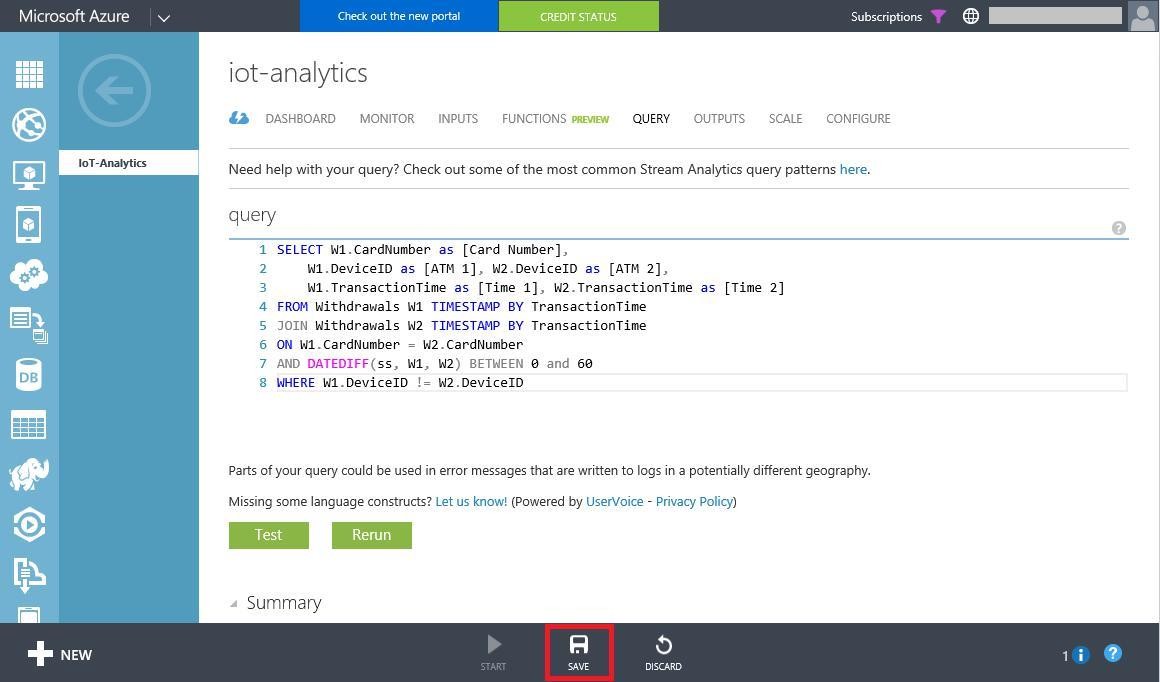


В этот раз вывод должен содержать только три строки, каждая из которых представляет две транзакции, выполненные с одной карточкой ATM в двух разных местах в течение 60 секунд друг от друга:



**Рис 4.7 Потенциально мошеннические транзакции**

Нажмите кнопку SAVE в нижней части страницы для сохранения запроса. Затем нажмите «YES», когда вас попросят подтвердить.



**Рис 4.8 Сохранение запроса**

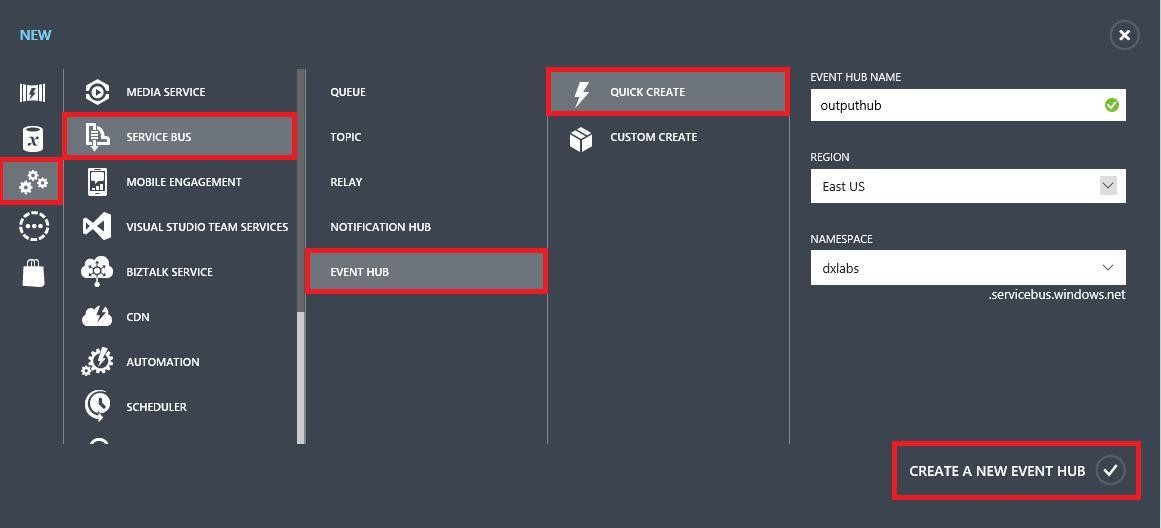
Теперь, когда запрос сформулирован, протестирован на основе набора выборочных данных и сохранен, пришло время развернуть его против реального потока данных, чтобы создать текущую запись потенциально мошеннических транзакций.

## Упражнение 5: Создание концентратора событий для вывода

Возможность запуска запросов и просмотра результатов на портале отлично подходит для тестирования, но все дело в том, что Stream Analytics может запрашивать потоки данных в реальном времени. Azure Stream Analytics поддерживает различные типы выходных данных, включая blobs, базы данных Azure SQL и концентраторы событий. Вы можете назначить хранилище больших двоичных объектов в качестве адресата для вывода из задания Stream Analytics для создания постоянной записи результатов запроса. Другой полезный сценарий предполагает использование концентраторов событий в качестве выходных данных. Поскольку программное обеспечение может подписаться на события из концентраторов событий, разработчики могут создавать собственные приложения, которые отображают вывод Stream Analytics в реальном времени.

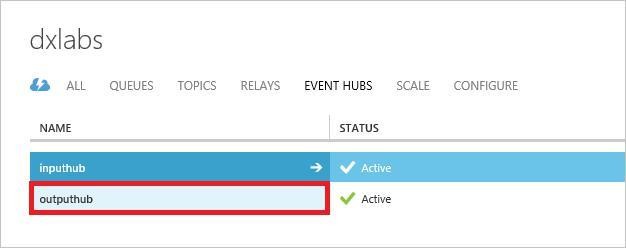
В этом упражнении вы настроите задание Stream Analytics на запись вывода в концентратор событий.

1. Вернитесь на классический портал и нажмите кнопку + NEW в левом нижнем углу.
2. Нажмите APP SERVICES -> SERVICE BUS -> EVENT HUB -> QUICK CREATE, чтобы создать новый концентратор событий. Назовите концентратор событий «outputhub», выберите тот же регион, который вы выбрали для другого концентратора событий и задания Stream Analytics, и выберите пространство имен, которое вы создали в упражнении 1. Затем нажмите «CREATE A NEW EVENT HUB»



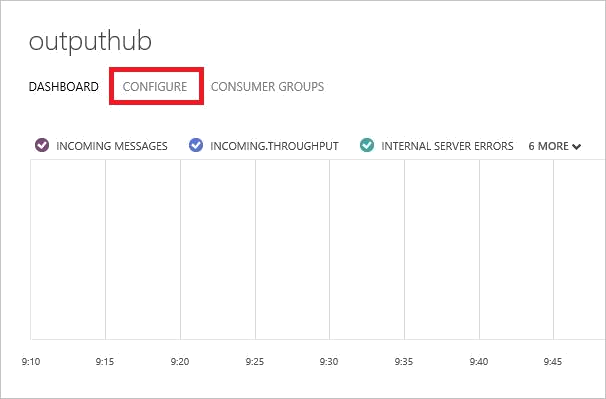
**Рис 5.1 Создание концентратора событий для вывода**

Нажмите новый концентратор событий, чтобы перейти на панель событийного концентратора.



**Рис 5.2 Открытие центра событий**

Нажмите CONFIGURE.

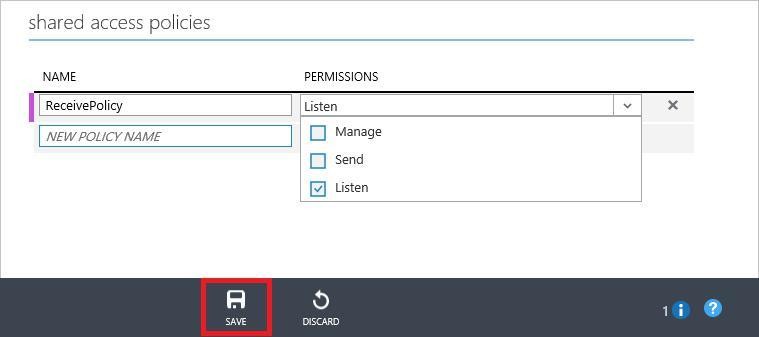


**Рис 5.3 Настройка концентратора событий**

Чтобы приложение могло подписаться на события, выходящие из концентратора событий, вам необходимо создать политику общего доступа, которая включает разрешение на прослушивание. В разделе политики общего доступа на странице конфигурации outputhub создайте новую политику, введя

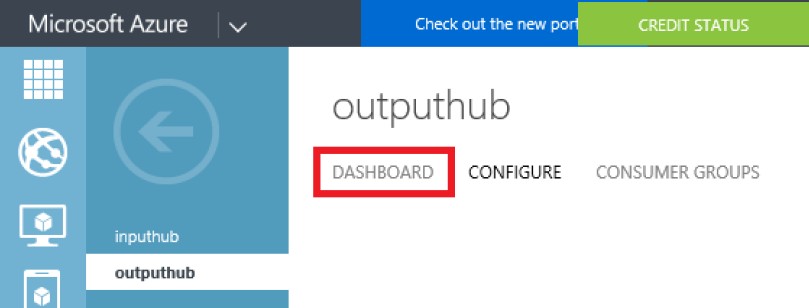
«ReceivePolicy» (без кавычек) в первое текстовое поле и проверив поле «Listen» в раскрывающемся списке в разделе «PERMISSIONS». Затем нажмите кнопку

«Save» внизу страницы, чтобы сохранить новую политику.



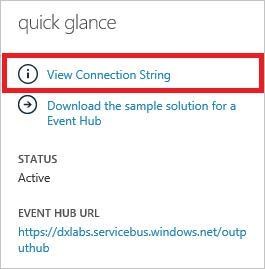
**Рис 5.4 Создание политики с разрешением на прослушивание**

Нажмите DASHBOARD в верхней части страницы, чтобы вернуться к панели инструментов мероприятия.



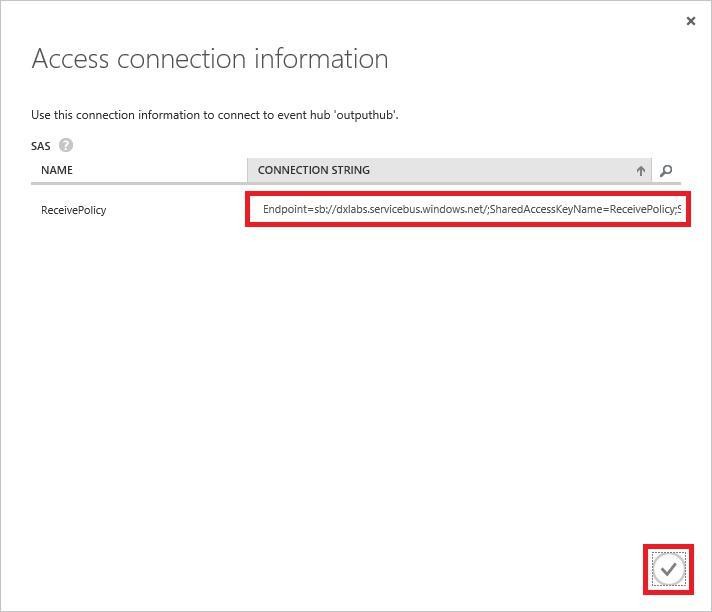
**Рис 5.5 Возвращение на панель инструментов**

Под quick glance на правой стороне страницы нажмите View Connection String.



**Рис 5.6 Просмотр строки подключения**

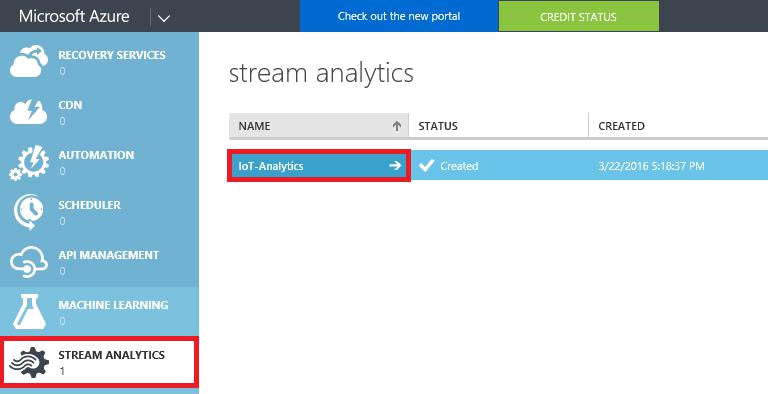
Скопируйте строку соединения в буфер обмена. (Если вы наведите курсор на строку соединения, появится кнопка копирования в буфер обмена.) Затем щелкните галочку, чтобы закрыть диалог. Завершите работу, вставив строку соединения в ваш любимый текстовый редактор, чтобы вы могли найти



ее в следующем упражнении.

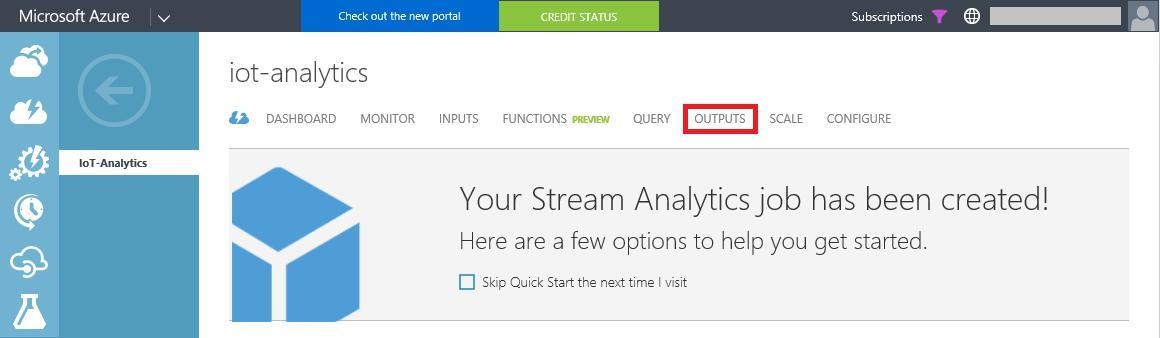
**Рис 5.7 Копирование строки подключения**

Нажмите STREAM ANALYTICS на ленте слева, чтобы перечислить все



задания Stram analytics, а затем нажмите IoT-Analytics, чтобы вернуться к заданию Stream Analytics.

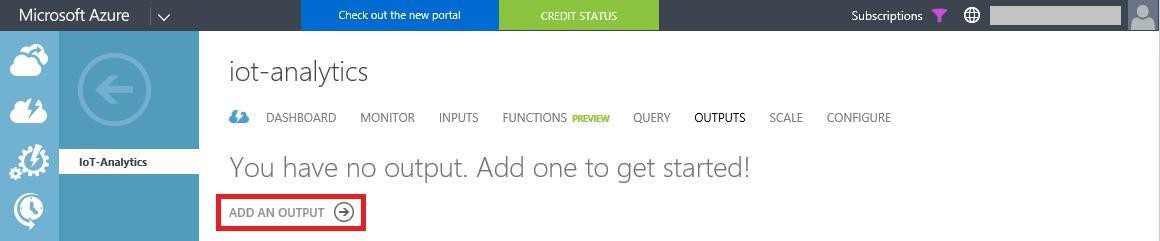
**Рис 5.8 Возврат к заданию Stream Analytics**



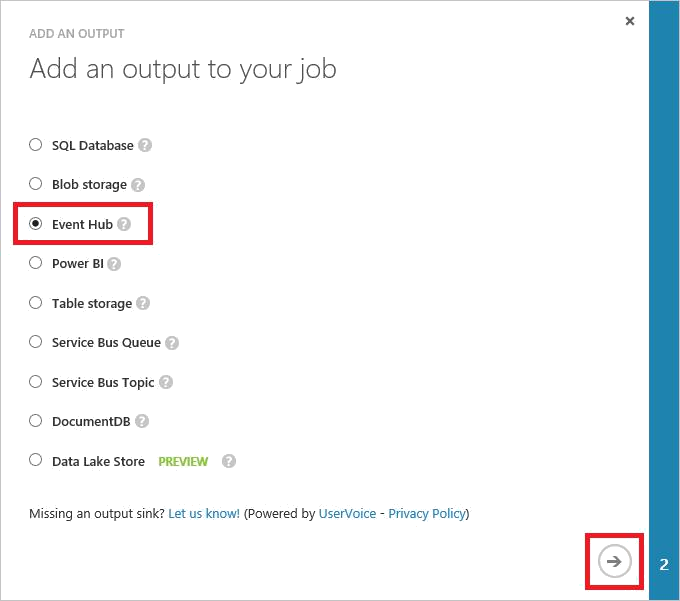
Нажмите OUTPUTS.

**Рис 5.9 Переход на страницу Outputs**

Нажмите ADD AN OUTPUT для добавления результата в задание.



**Рис 5.10 Добавление результата**

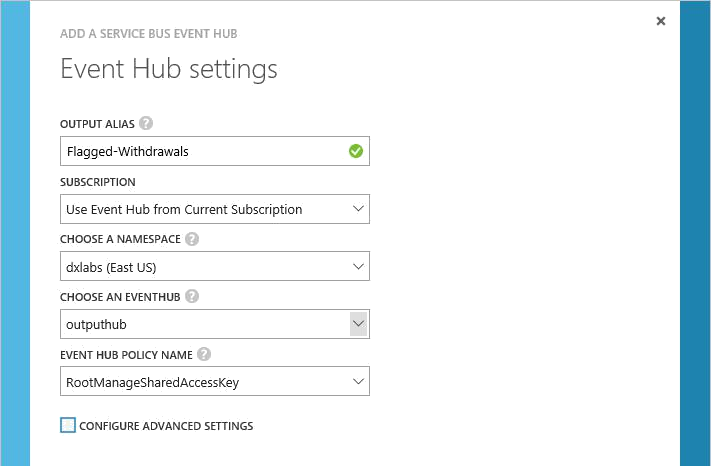


Выберите Event Hub как тип вывода. Затем нажмите правую стрелку в нижнем правом углу диалога.

**Рис 5.11 Указание типа вывода**

Введите «Flagged-Withdrawals» в поле OUTPUT ALIAS. Выберите

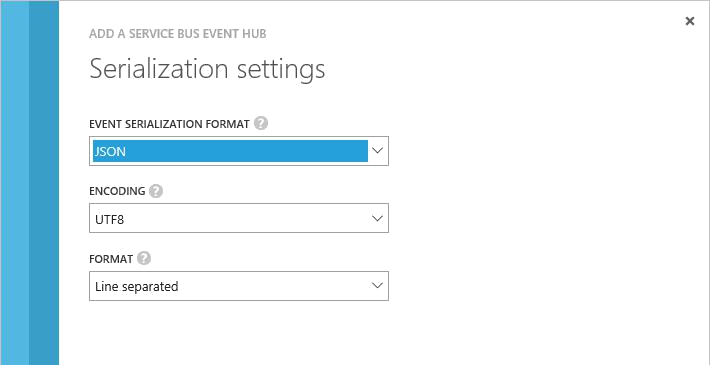
«Использовать Use Event Hub from Current Subscription» из текущей подписки, пространство имен, созданное в упражнении 1, и концентратор событий, созданный в этом упражнении (outputhub). Затем нажмите правую стрелку в нижнем правом углу.



**Рис 5.12 Задание настроек центра событий**

Убедитесь, что EVENT SERIALIZATION FORMAT, ENCODING и

FORMAT установлены, как показано ниже, а затем закончите, щелкнув галочку в нижнем правом углу.



**Рис 5.13 Указание настроек сериализации**

Теперь, когда вы направили вывод из задания Stream Analytics на концентратор событий, следующая задача - написать приложение, которое потребляет события из этого концентратора событий.

## Упражнение 6: Создание панели управления в реальном времени

В этом упражнении вы напишете веб-приложение, которое подключается к концентратору событий и отображает уведомления о потенциально мошеннических транзакциях в режиме реального времени.

Запустите новый экземпляр Visual Studio 2015 и используйте команду File

-> New -> Project чтобы создать новое веб-приложение ASP.NET с именем

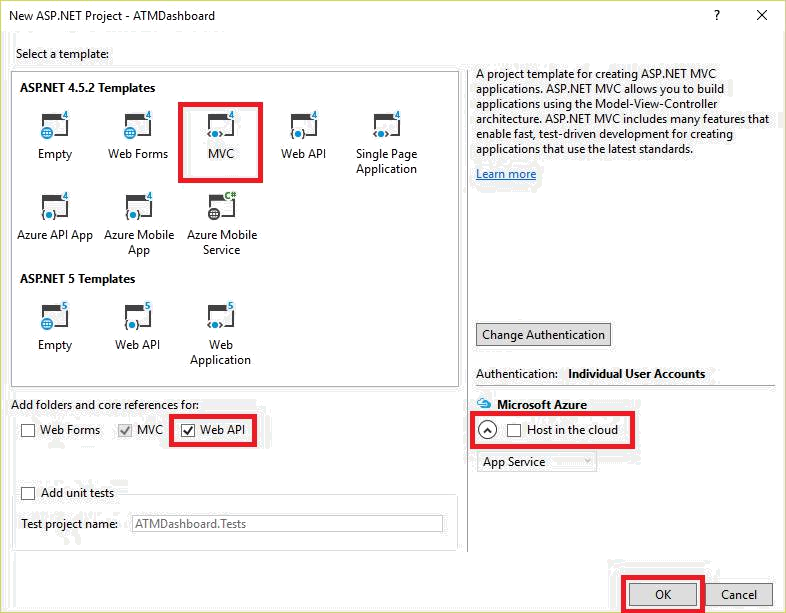
«ATMDashboard."



**Рис 6.1 Создание нового веб-приложения**

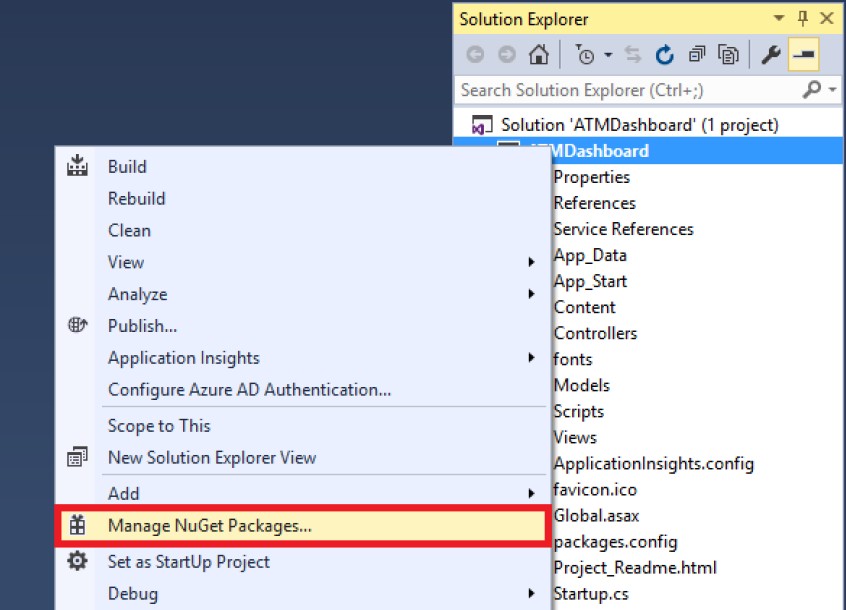
В диалоговом окне «New ASP.NET Project» выберите MVC и установите флажок «Web API». Если хост в облаке установлен, снимите его. (Для тестирования вы запустите это веб-приложение локально.) Затем нажмите

«OK».



**Рис 6.2 Определение параметров для веб-приложения**

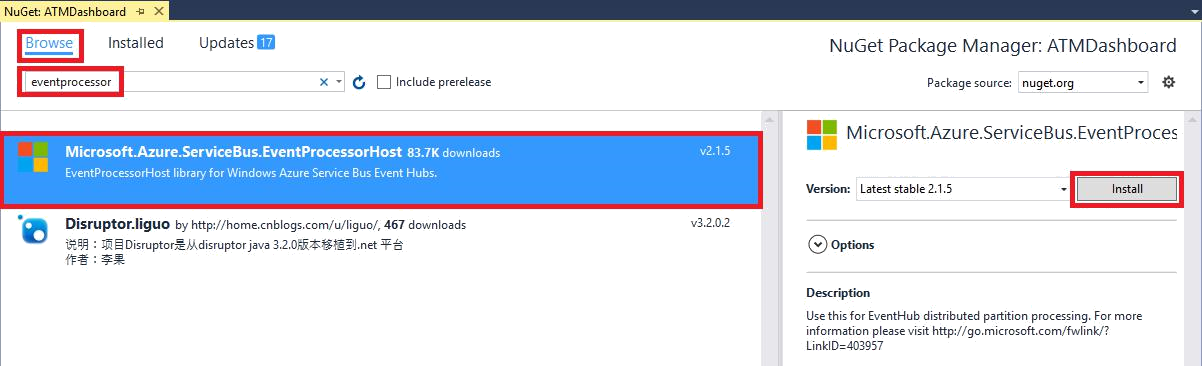
В окне «Solution explorer» щелкните правой кнопкой мыши проект



ATMDashboard и выберите «Manage NuGet Packages».

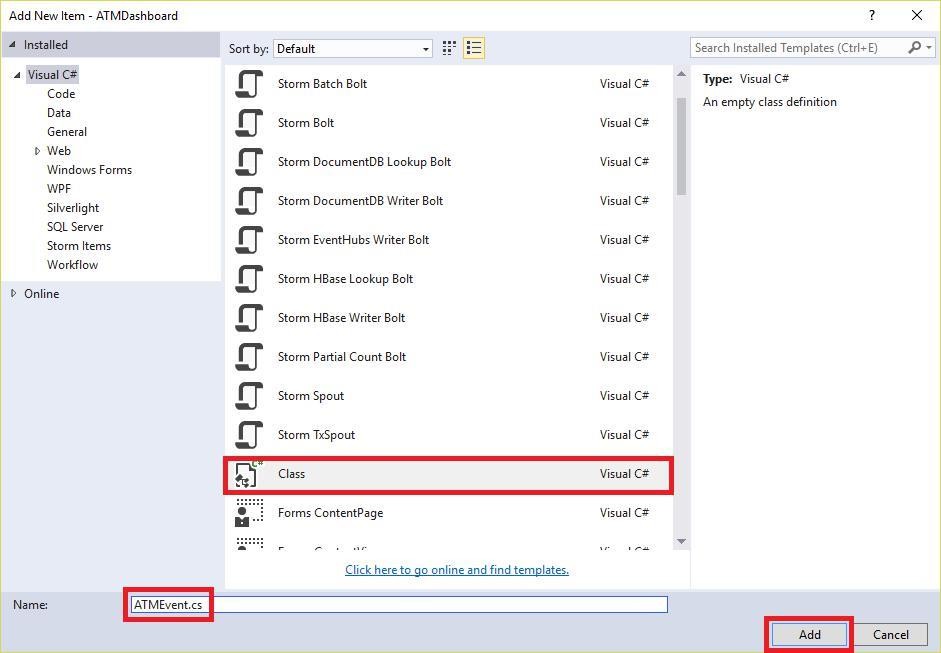
**Рис 6.3 Управление пакетами NuGet для проекта**

Наконец, нажмите «Install», чтобы установить последнюю стабильную версию пакета. Этот пакет содержит API-интерфейсы, которые ваше приложение будет использовать для получения событий из выходного концентратора событий. Нажмите OK, если вам будет предложено просмотреть изменения, и I Accept, когда появится приглашение принять лицензии для загруженных пакетов.



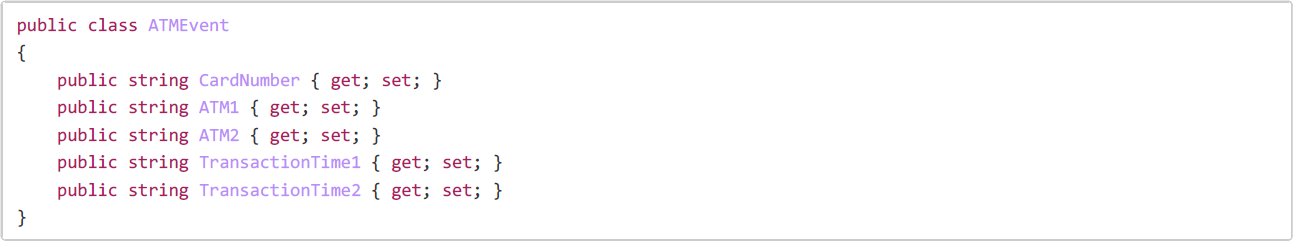
**Рис 6.4 Установка EventProcessorHost**

Щелкните правой кнопкой мыши проект в окне Solution Explorer и используйте команду Add -> Class, чтобы добавить к проекту класс ATMEvent.

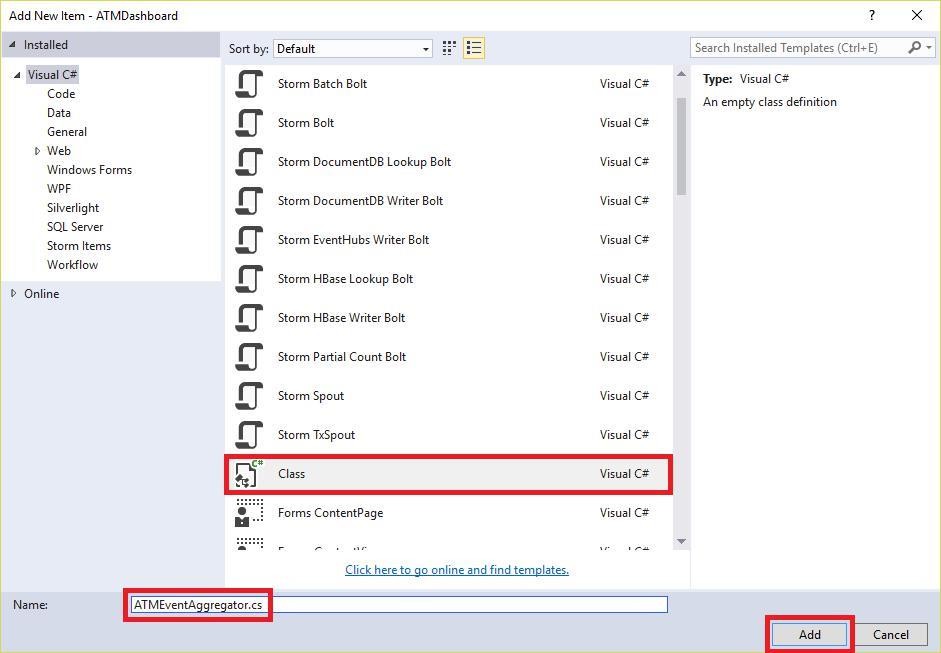


**Рис 6.5 Добавление ATMEvent**

Внедрите класс ATMEvent следующим образом:

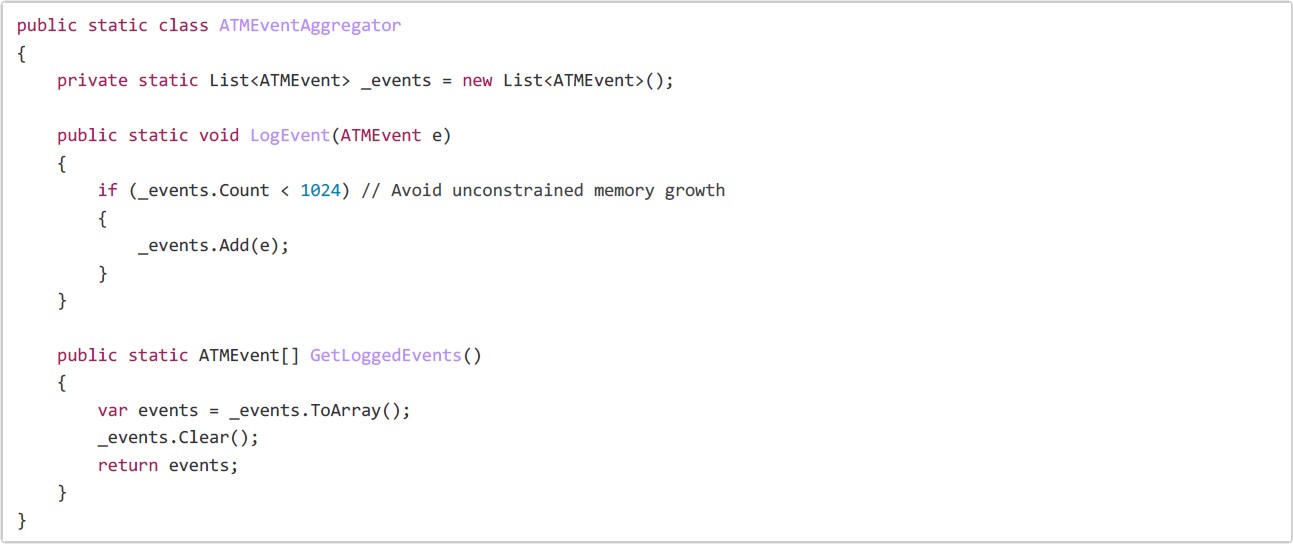


Щелкните правой кнопкой мыши проект в окне Solution Explorer и используйте команду Add -> Class, чтобы добавить в проект класс ATMEventAggregator.

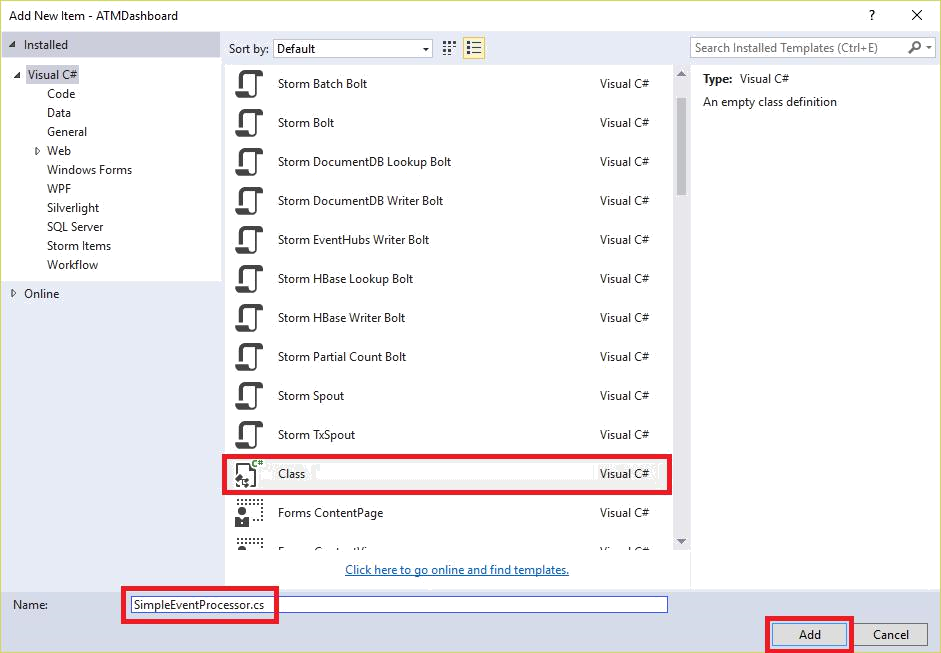


**Рис 6.6 Добавление ATMEventAggregator**

Внедрите класс ATMEventAggregator следующим образом:



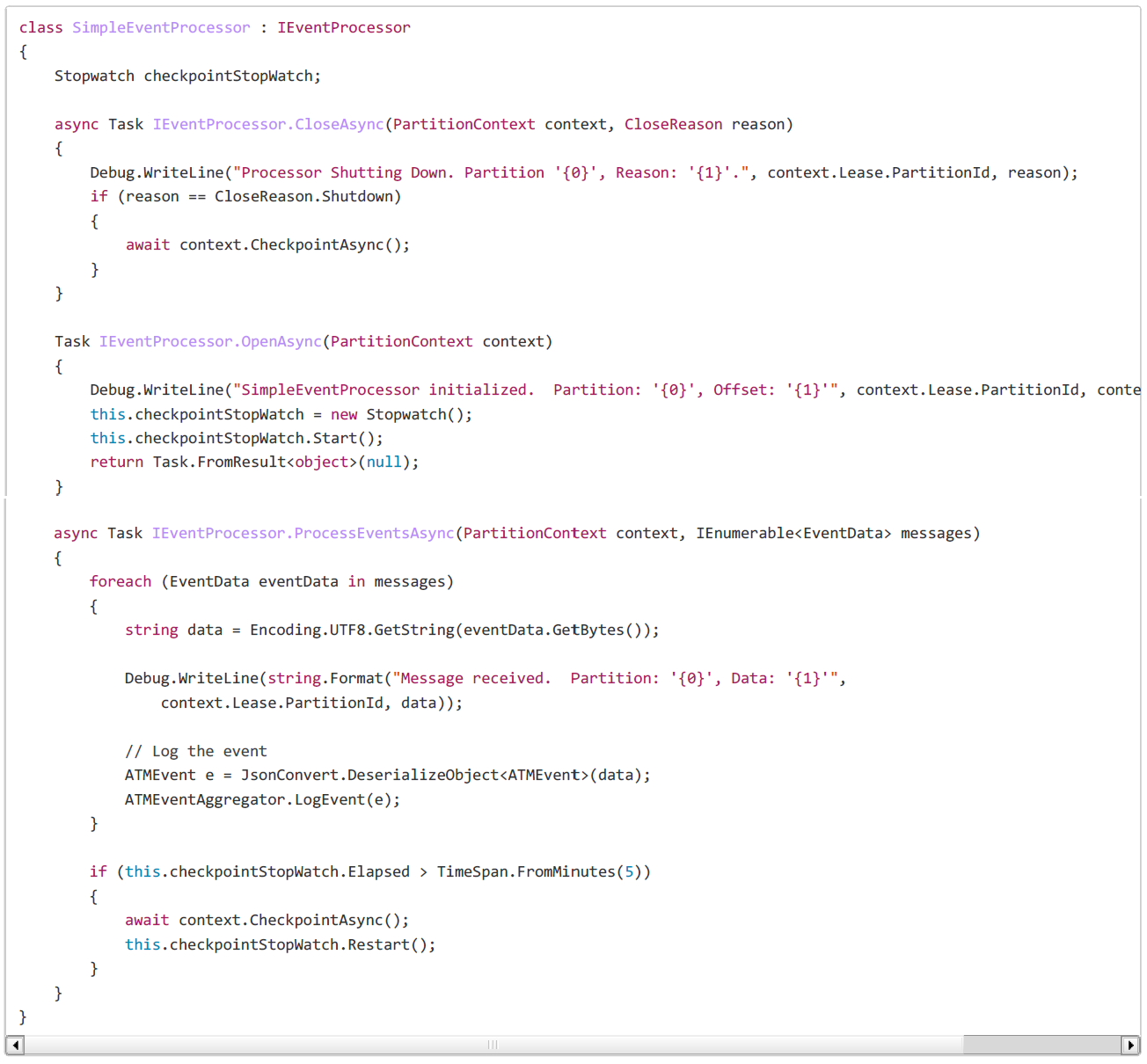
Щелкните правой кнопкой мыши проект в окне Solution Explorer и используйте команду Add -> Class, чтобы добавить к проекту класс



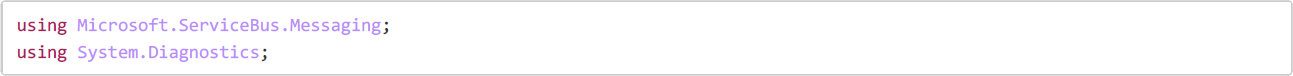
SimpleEventProcessor.

**Рис 6.7 Добавление SimpleEventProcessor**

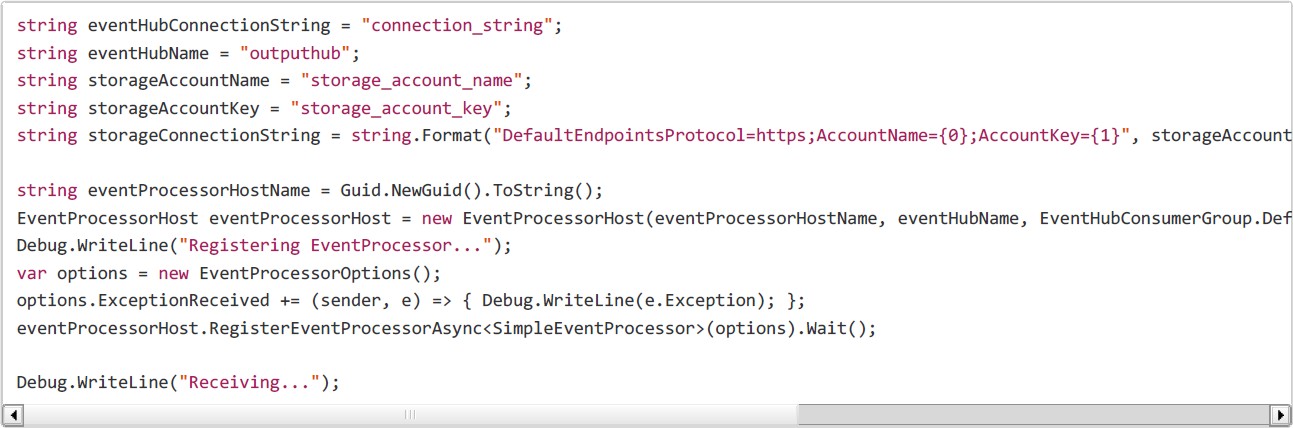
Добавьте следующие операторы using к тем, которые расположены вверху файла:



Откройте Global.asax.cs и лобавьте следующие операторы using:



В Global.asax.cs, добавьте следующие операторы в конец метода Application\_Start:



Замените connection\_string в строке 1 с помощью строки подключения, которую вы скопировали в буфер обмена в упражнении 5, шаг 8.

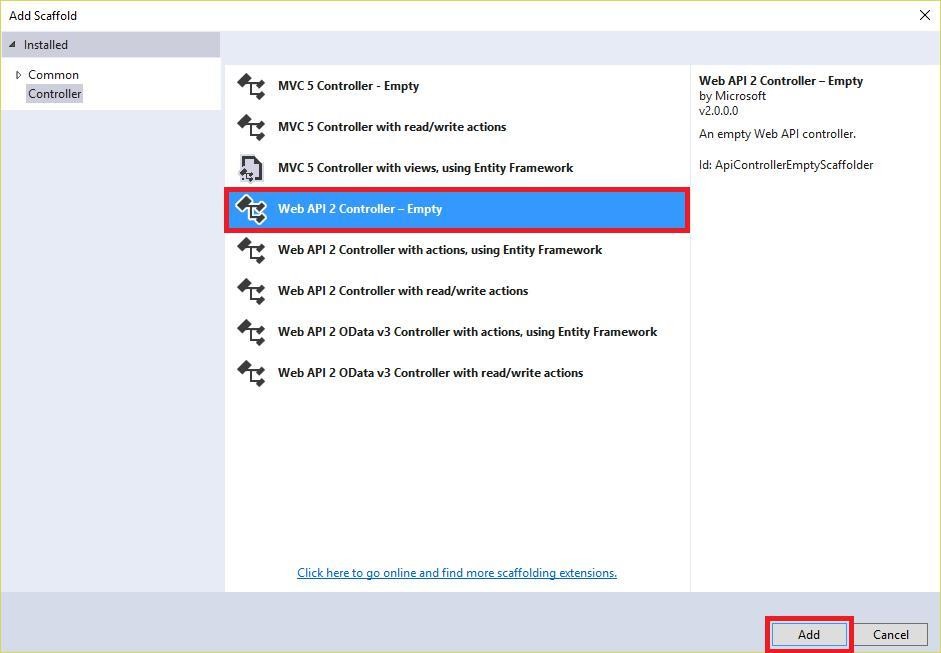
Удалите ";EntityPath=outputhub" из конца строки подключения. Объявление переменной eventHubConnectionString должно теперь выглядеть примерно так::

Замените storage\_account\_name в строке 3 с именем storage account, созданным вами в упражнении 3, шаг 2, и файле storage\_account\_key в строке 4 с основным ключом учетной записи хранилища. Вы можете скопировать ключ доступа в буфер обмена, открыв учетную запись хранилища на портале, нажав кнопку «MANAGE ACCESS KEYS в нижней части страницы и нажав кнопку

«Copy» рядом с «to PRIMARYACCESS KEY»

Эта учетная запись хранилища не имеет ничего общего с заданием Stream Analytics; Он используется классом EventProcessorHost. Тем не менее, вы позволяете одной учетной записи хранилища делать двойную нагрузку для Stream Analytics и EventProcessorHost.

В окне Solution Explorer щелкните правой кнопкой мыши папку Controllers и используйте команду Add -> Controller, чтобы добавить пустой Web API 2 контроллер.



**Рис 6.8 Добавление контроллера веб-интерфейса API**

Назовите контроллер «EventsController» (без кавычек). Затем нажмите



«Add».

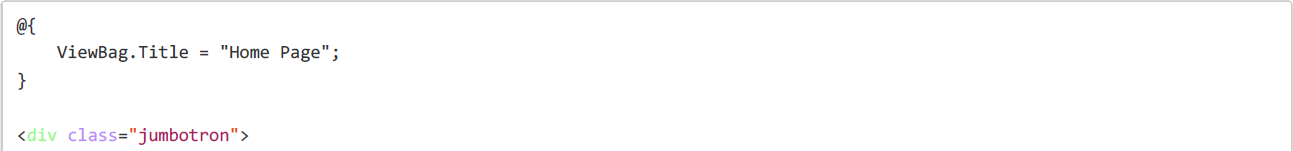
**Рис 6.9 Именование контроллера веб-интерфейса API**

Добавьте следующий класс в класс EventsController.



Контроллер, который вы только что добавили, является контроллером веб-интерфейса API. Он предоставляет способ REST, который можно вызывать через HTTP для извлечения последних событий, полученных ATMEventAggregator от концентратора событий.

Откройте Index.cshtml в папке Views / Home проекта и замените его содержимым на следующие операторы:



Видите, что здесь происходит? В дополнение к изменению пользовательского интерфейса представления для включения таблицы HTML, в которой могут отображаться события ATM, вы добавили блок сценария, который использует $ .ajax метод jQuery для обратного вызова на сервер каждые 5 секунд. Конечной точкой для вызова является метод, реализованный в контроллере веб-интерфейса API на предыдущем шаге. Когда этот метод возвращает одно или несколько событий, строки добавляются в таблицу для их отображения.

Перейдите в меню «Build» в верхней части окна Visual Studio и используйте команду «Build Solution», чтобы создать решение. Исправьте все



сообщения об ошибках сборки, а затем нажмите Ctrl+F5, чтобы запустить приложение в своем браузере. Убедитесь, что приложение выглядит так:

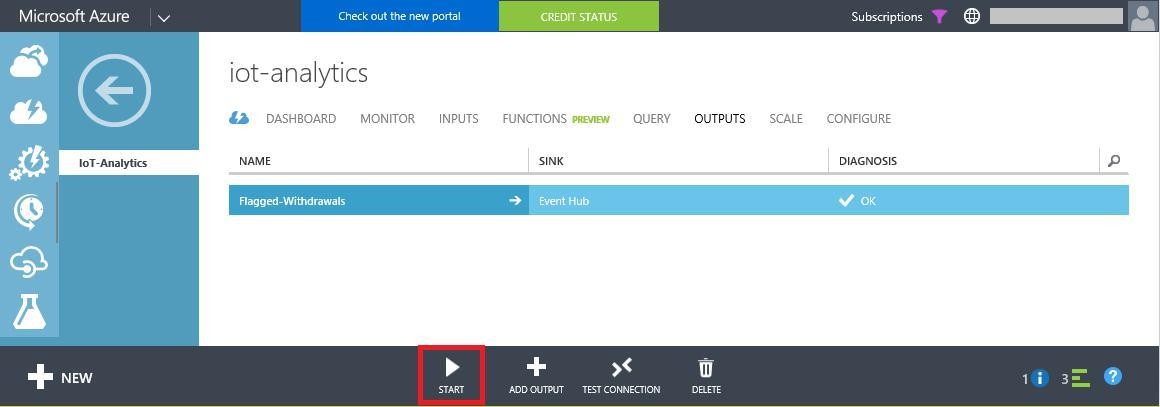
**Рис 6.10 Панель управления банкоматом**

Оставьте браузер открытым при запущенном приложении. И не пугайтесь, если несколько событий появляются в первые несколько секунд, когда панель инструментов открыта. В следующем упражнении вы будете создавать новые события.

## Упражнение 7: Анализ потока данных в реальном времени

Почти готово! Теперь пришло время запустить задание Stream Analytics и увидеть результат на панели инструментов.

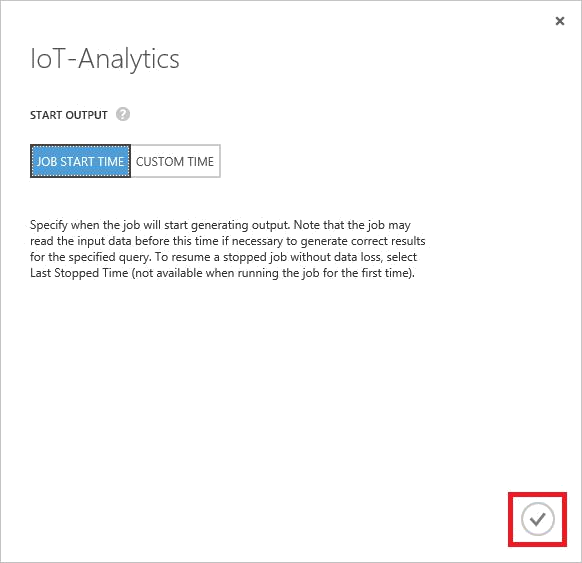
Вернитесь к заданию Stream Analytics на портале Azure и нажмите кнопку



START в нижней части страницы, чтобы запустить его.

**Рис 7.1 Запуск задания Stream Analytics**

Убедитесь, что JOB START TIME выбрано, а затем щелкните галочку. JOB START TIME (ВРЕМЯ ЗАПУСКА ЗАДАНИЯ) означает, что задание начнет формировать выборку с входного источника в момент запуска задания. Концентраторы событий сохраняют события в течение заданного периода



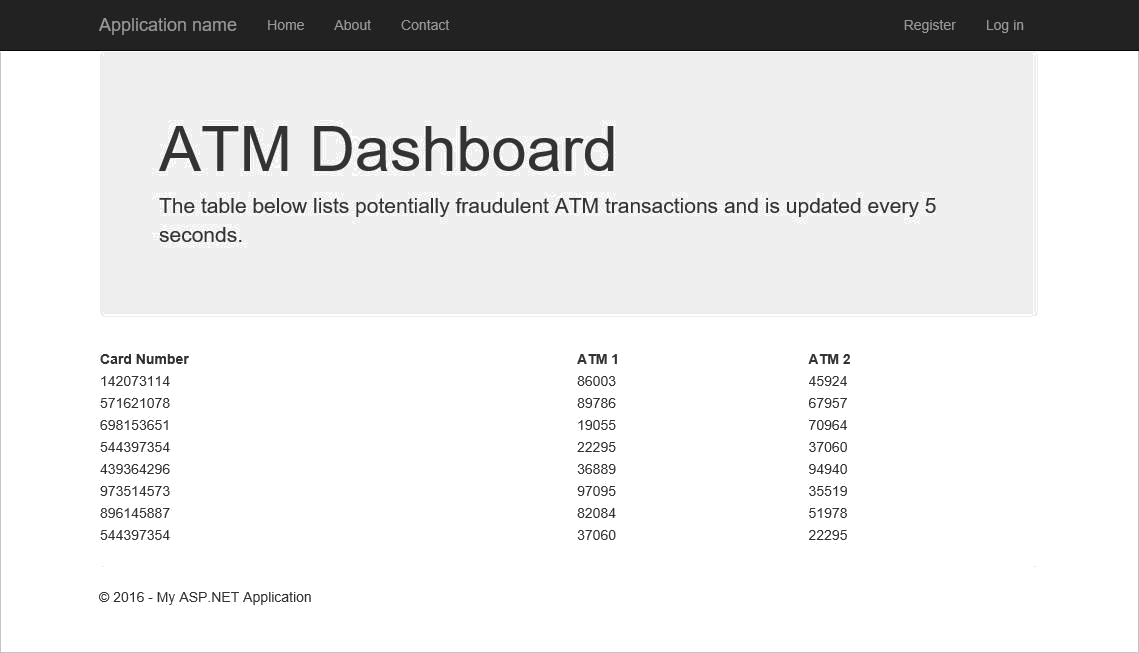
времени (по умолчанию 1 день), поэтому, если вы хотите, вы можете выбрать CUSTOM TIME и начать выборку до начала работы.

**Рис 7.2 Запуск вывода**

Дождитесь запуска задания Stream Analytics. Затем вернитесь к экземпляру Visual Studio, в котором открыт проект ATMEventGenerator, и запустите ATMEventGenerator, чтобы перенести события во входной концентратор событий.

Если вы хотите, вы можете запустить несколько экземпляров ATMEventGenerator одновременно, чтобы увеличить объем событий.

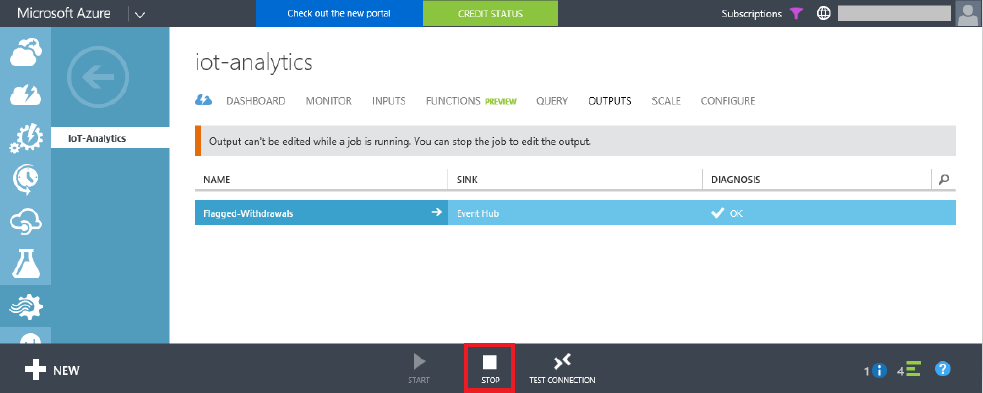
Вернитесь к приложению ATMDashboard, запущенному в вашем браузере, и смотрите его несколько минут. Время от времени должна появляться строка, представляющая потенциально мошенническую транзакцию, и идентификационный номер банкомата и банкоматы, в которых использовалась



карта.

**Рис 7.3 Потенциально мошеннические транзакции**

Вернитесь к заданию Stream Analytics на портале и нажмите кнопку STOP в нижней части страницы, чтобы остановить его. Затем нажмите YES, если вы уверены, что хотите остановить работу.



**Рис 7.4 Остановка задания Stream Analytics**

Вернитесь в консольное окно, в котором запущен ATMEventGenerator, и нажмите Ctrl + C, чтобы прервать его.

Теперь у вас есть веб-приложение, отображающее вывод задания Stream Analytics в режиме, близком к реальному. Существуют и другие способы создания таких панелей мониторинга, в том числе Microsoft Power BI. С помощью Power BI вы можете создавать информационные панели, которые выводят выходные данные из заданий Stream Analytics без написания кода. Дополнительные сведения см. В разделе Аналитика потоков и Power BI. Панель мониторинга аналитики в реальном времени для потоковой передачи данных.

# Заключение

Azure Stream Analytics - это мощный инструмент для анализа потоков живых данных с устройств IoT или всего остального, способного передавать данные. В этой лаборатории, вы получили непосредственный взгляд на Stream Analytics, а также на концентраторы событий Azure. Помимо прочего, вы узнали, как:

* + Создайте концентраторы событий Azure и используйте их для ввода и вывода Stream Analytics
  + Напишите приложение C #, которое передает события в концентратор событий
  + Создайте задание Stream Analytics и проверьте запросы на выборочные потоки данных
  + Запустите задание Stream Analytics и выполните запросы в потоках реальных данных
  + Создайте правило (запрос), которое обнаруживает аномалии в потоковых данных
  + Отображать вывод из задания Stream Analytics в веб-приложении

Один из недостатков правил жесткого кодирования в Stream Analytics заключается в том, что правила не «учатся» из потоков данных, что может привести к ложным срабатываниям в обнаружении аномалий. Azure Machine учится на представленных им данных. Представьте, что объединение Stream Analytics позволяет извлекать информацию из реального времени потока данных с мощью Azure Machine Learning, чтобы учиться на этой информации и совершенствовать аналитику «на лету». Это как раз тип решения, которое Microsoft Azure дает вам возможность создавать!