УДК 004

ИНСТРУМЕНТАРИИ WINDOWS AZURE ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

А.И. Трефилова
Руководитель: А.А. Алексеев, аспирант кафедры ОСУ ИК ТПУ
Томский Политехнический Университет
E-mail: alyona.trefilova@gmail.com

This article describes Windows Azure's solutions for Data Mining.

Key words: clouds, cloud platform, Data Mining, Windows Azure.
Ключевые слова: облака, облачная платформа, анализ данных, Windows Azure.

Введение

Интеллектуальный анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных. [1] Облачная платформа Windows Azure предоставляет пользователю различные сервисы, службы и инструменты для обеспечения такого анализа. Облачные платформы Amazon и Google также предоставляют сервисы для обработки и анализа данных, но, тем не менее, Windows Azure обладает рядом преимуществ перед этими платформами. Amazon Web Services предлагает использовать только одно облако, а у Windows Azure есть гибридное облако, использующее ресурсы площадки заказчика и ресурсы облака. Также дата-центры Azure охватывает более широкие пространства, чем Google и Amazon. На рис. 1 представлена схема сервисов Windows Azure для анализа данных. В рамках данной статьи будет рассмотрен список задач решаемых средствами Windows Azure.

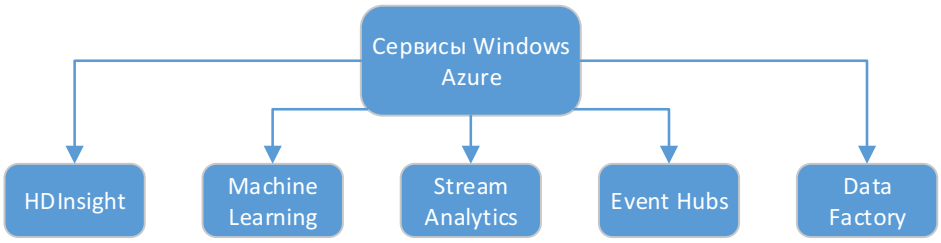

Рис. 1. Сервисы Windows Azure для анализа данных

HDInsight

HDInsight – это решение, позволяющее работать с программной платформой Apache Hadoop и построенное на основе облачных технологий. Данное средство позволяет обработать большой объем информации с масштабированием по мере необходимости. Также HDInsight работает как со структурированными данными, так и с полуструктурированными и неструктурированными. [2] Данное решение применяется для задач поиска и классификации, типовых задач обработки данных, задач управления потоками данных, запросных, аналитических задач. Например, выявление сложных зависимостей в социальных сетях, антифрод-системы в финансовом секторе, исследование генома, анализ логов и т. д.

Фабрика данных (Data Factory)

Фабрика данных – это управляемая служба для хранения и обработки данных, работающая как со структурированными, так и неструктурированными данными, полученными как из локальных, так и из облачных источников. Фабрика данных используется для подключения расположенных в разных источниках данных различных типов, ввода в эксплуатацию конвейеров данных, интеграции производства информации с обнаружением данных. [3]

Машинное обучение (Machine Learning)

Machine Learning – это сервис машинного обучения в облаке, который позволяет решать задачи аналитики с прогнозированием. Сервис состоит из двух компонентов: Machine Learning Studio (клиентская часть) и Machine Learning API Service (серверная часть). [4] Сервис Machine Learning нашел свое применение в кластеризации методом k-средних, классификации методом ближайшего соседа, прогнозировании степени злокачественности рака молочной железы, анализе тональности текста, обработке некоторого входного текста, извлечения из него именованных терминов и их автоматической классификации по категориям (к примеру, Люди или Места).

Концентраторы событий (Event Hubs)

Концентраторы событий – это служба для обеспечения приема событий и телеметрических данных с высокой степенью масштабируемости. Данный сервис позволяет собирать телеметрию приложений, проводить диагностику данных, устройств телеметрии [5].

Потоковый обработчик событий (Stream Analytics)

Stream Analytics – это сервис для анализа потока данных в реальном времени. Для получения потока данных взаимодействует с Azure Event Hub и хранилищем, а для хранения результатов анализа – с Event Hubs, Blob Storage, Azure SQL Database. [6] Данный сервис в комплексе с концентраторами событий (Event Hubs) дает возможность обрабатывать большие объемы данных в режиме реального времени. Сервис можно использовать для персонализированного торгового анализа в режиме реального времени, выявления мошенничества в режиме реального времени, служб защиты данных и личной информации, сбора и анализа данных, поступающих от датчиков, потоковой аналитики переходов на сайтах и CRM-приложений, отправляющих сигналы при падении уровня обслуживания клиента в течение определенного промежутка времени.

Заключение

Анализ и обработка большого массива данных при помощи облачных сервисов и служб позволяет максимально быстро и качественно получать результаты в сферах, где пользователям приходится работать с большими объемами неструктурированной информации, на основании которой принимаются стратегические решения. Облачная платформа Windows Azure в значительной степени позволяет упростить эти процессы. Анализ данных в режиме реального времени, масштабируемость по запросу пользователя, автоматизированное управление и т. д. – это лишь малый спектр задач, решаемые в рамках этой платформы.

В ходе анализа списка задач, решаемых платформой Microsoft Windows Azure, было принято решение использовать машинное обучение (Machine Leaning) в рамках магистерской диссертационной работы, где планируется использовать не только разработки собственного программного обеспечения для кластерного анализа данных, но и применить для исследований облачные решения.

Список литературы

1. Основные понятия интеллектуального анализа данных [Электронный ресурс]. – Режим доступа: https://msdn.microsoft.com/ru-ru/library/ms174949.aspx, свободный.

2. Облачные службы – HDInsight (Hadoop) [Электронный ресурс]. – Режим доступа: http://azure.microsoft.com/ru-ru/services/hdinsight/, свободный.

3. The Ins and Outs of Azure Data Factory – Orchestration and Management of Diverse Data [Электронный ресурс]. – Режим доступа: http://blogs.technet.com/b/dataplatforminsider/archive/2014/10/30/the-ins-and-outs-of-azure-data-factory-orchestration-and-management-of-diverse-data.aspx, свободный.

4. Alex Belotserkovskiy – Microsoft представляет новый сервис машинного обучения Azure Machine Learning [Электронный ресурс]. – Режим доступа: http://blogs.msdn.com/b/albe/archive/2014/07/15/microsoft-azure-machine-learning.aspx, свободный.

5. Облачные службы – концентраторы событий [Электронный ресурс]. – Режим доступа: http://azure.microsoft.com/ru-ru/services/event-hubs/, свободный.

6. Azure Newsletter [Электронный ресурс]. – Режим доступа: http://ibmpw.blogspot.ru/2015/03/azure-newsletter-2015.html, свободный.

УДК 004

КЛАСТЕРИЗАЦИЯ СПУТНИКОВЫХ СНИМКОВ С ИСПОЛЬЗОВАНИЕМ ОБЛАЧНЫХ ТЕХНОЛОГИЙ MACHINE LEARNING AZURE

Е.Е. Васильева
Научный руководитель: Н.Г. Марков, д.т.н., профессор, зав.каф. ВТ ИК ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: ekaterina.vasilyeva9@gmail.com

This article deals with a modern cloud computing technology – Machine Learning Azure. The technology gives a variety of tools to design the systems which implement popular machine learning algorithms and approaches such as anomaly detection, classification, clustering and regression. It can be used for data analysis in different fields but article shows how it can be used for satellite image processing and recognition.

Keywords: cloud computing, clustering, satellite image recognition, Machine Learning Azure.
Ключевые слова: облачные технологии, кластеризация, распознавание спутниковых снимков, Machine Learning Azure.

Введение

В связи с развитием методов машинного обучения и компьютеризацией, все большее количество задач становится возможным решать при их помощи. В связи с этим Microsoft представили в 2014 г. новый облачный сервис – Machine Learning Azure [1], позволяющий пользователям по мере необходимости быстро анализировать данные. Это графическая среда, в которой некоторые инструменты и алгоритмы машинного обучения представлены в виде блоков. Для реализации собственных алгоритмов используются блоки, компилирующие код на языке R или Python. Создавая определенную схему из блоков с корректно настроенными параметрами, можно получить модели классификации, кластеризации, линейной регрессии и статистического анализа. Далее рассмотрена кластеризация спутникового снимка при помощи Machine Learning Azure.

Обработка данных

В ходе исследования использовался набор данных Urban Land Cover Dataset [2], который состоит из текстур, описываемых множеством признаков (яркость, индекс формы, NDVI (вегетационный индекс), средние значения в красном и зеленом каналах и др.), полученных из космоснимка высокого разрешения. Эти данные нуждаются в предварительной обработке. К мерам обработки относятся: удаление дубликатов, восстановление отсутствующих значений, нормализация.

Нормализация данных необходима, когда параметры измеряются в разных шкалах. Рассматриваемый набор данных был нормализирован по Z-score: