ОБРАБОТКА БОЛЬШИХ ОБЪЕМОВ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ HADOOP И СЛУЖБЫ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ MICROSOFT AZURE

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Хадасевич А.И., Швец В.И.

Хотеев А.Л. – к.ф.-м.н., доцент

С каждым годом задачи обработки больших объемов данных все чаще становятся перед разработчиками. В настоящее время генерируются данные о деятельности людей и объектов в огромном количестве и растущем масштабе. Для изучения, поиска закономерностей и анализа этих данных используются специальные инструменты и методы.

Платформа Hadoop представляет собой набор утилит, библиотек и фреймворков для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов с использованием простых моделей программирования. Он также масштабируемый от отдельных серверов до тысяч компьютеров, каждый из которых предлагает локальные вычисления и хранение. Данная платформа является своего рода стандартом для решения задач, связанных с полу- и неструктурированными большими данными. Однако для начала работы с платформой Hadoop, требуется много времени для изучения основ платформы. Также достаточно велики начальные финансовые затраты на аппаратное обеспечение такого рода кластера. Эти причины так же усложняют обучение для работы с данной платформой. Одно из возможных решений приведенных проблем, является объединение облачных технологий и платформы Hadoop. Это достаточно многообещающий способ решения проблемы «Больших данных», имеющий низкий уровень входа (квалификация разработчика и стоимость на запуск).

Компания Microsoft предлагает Windows Azure HDInsight — это полностью управляемая облачная служба, которая позволяет быстро, просто и без лишних затрат обрабатывать большие объемы данных. Данное облачное решение поддерживает различные сценарии:

1) хранение данных;
2) обработка, анализ данных и машинное обучение;
3) извлечение, преобразование и загрузка;
4) "интернет вещей";
5) приложения потоковой передачи.

Для разработчиков Windows Azure HDInsight имеет плагин для Visual Studio, который поддерживает создание приложений. Для разработчиков Linux или Windows у HDInsight есть плагины для IntelliJ IDEA и Eclipse, двух очень популярных платформ Java IDE с открытым исходным кодом. HDInsight также поддерживает команды PowerShell, Bash и Windows, позволяющие создавать сценарии рабочих процессов. Также отличительной особенностью является возможность использования не JVM языков с платформой Hadoop. Таким образом увеличивается количество разработчиков, которые могут работать с Hadoop. JavaScript API позволяет разработчикам определять задачи Hadoop в более естественном стиле. Разработчики .NET также могут работать с Hadoop, используя преимущества языковых функций C# и F#, таких как LINQ.

Для ученых-исследователей, HDInsight включает Jupyter. Jupyter- это среда разработки на базе обычного текстового редактора, которая позволяет интегрировать код и контент. Когда код и контент объединяются, они создают живой документ, который обновляется данными.

Также компания Microsoft предоставляет удобную бесплатную среду разработки WindowsAzureHDInsightEmulator. Это реализация HDInsight для Windows, предоставляющая набор служб на базе Apache Hadoop в Windows используя только одноуровневое развертывание. Сервер HDInsight обеспечивает локальную среду разработки для службы Windows Azure HDInsight. Эта технология разрабатывается для обеспечения программной среды, предназначенной для управления, анализа и отчетности по большим данным. Подобно службе Windows Azure HDInsight, эта локальная среда разработки для HDInsight упрощает настройку, выполнение и пост-обработку заданий Hadoop, предоставляя библиотеку PowerShell с командлетами HDInsight для управления кластером и выполняемыми на нем заданиями.

Windows Azure HDInsight является гибким решением, которое позволяет программистам за небольшое время развернуть Hadoop кластер или выбрать требуемую услугу в облаке. Windows Azure HDInsight снижает уровень входа для разработчиков, которые будут работать с приложениями, связанными с обработкой больших объемов данных. Разработчики, которые ранее не работали с большими данными, получают большой выбор языков программирования и знакомые инструменты разработки.

Список использованных источников:

1. AzureHDInsightDocumentation. [Электронный ресурс]. – Электронные данные. – Режим доступа: https://docs.microsoft.com/en-us/azure/hdinsight/

2. HDInsight - Fully managed, full spectrum open-source analytics service for enterprises. [Электронный ресурс]. – Электронные данные. – Режим доступа: https://azure.microsoft.com/en-us/services/hdinsight/