ГЕОДЕЗИЯ

УДК 528.065/.067:004.6

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ОБРАБОТКИ БОЛЬШИХ ГЕОПРОСТРАНСТВЕННЫХ ДАННЫХ

К.С. АЛЕКСЕЕВА, А.В. КИРИЛЛОВА
(Представлено: П.Ф. Парадня)

Статья посвящена анализу программного обеспечения для обработки больших данных (Big Date). Особое внимание уделено программному продукту Google BigQuery GIS, как наиболее подходящему для целей хранения и обработки больших объемов геопространственной информации. Выделены основные характеристики и функции BigQuery. Автором написан скрипт на языке Python для получения геоинформации с сайта Невадской геодезической лаборатории, фрагмент которого и описание приведены в статье.

Мы живем в быстро меняющемся мире цифровых данных и информации. Каждую секунду многочисленные датчики и средства цифровой связи собирают гигабайты и терабайты информации с нашей помощью или независимо от нас обо всем, что нас окружает. Эра больших данных уже началась, и остановить ее уже невозможно, поскольку это будет означать конец научно-технического прогресса. Нет точного определения термина «большие данные». Первоначально идея заключалась в том, что объем информации стал настолько большим, что он фактически не помещался в память компьютера, используемого для обработки, и инженерам пришлось модернизировать инструменты для анализа всех данных. Так появились новые технологии обработки данных (такие как модель Google MapReduce и ее аналог с открытым исходным кодом, Hadoop от Yahoo). Они позволили управлять гораздо большим количеством данных, чем раньше. Кроме того, появились и другие технологии обработки данных, которые ранее также обходились без строгой иерархии и единообразия [1].

На данный момент существует большое количество инструментов Big Data для анализа данных. Анализ данных представляет собой процесс проверки, очищения, преобразования данных для получения необходимой информации. Имеются Big Data-инструменты с открытым исходным кодом, инструменты извлечения, визуализации, баз данных и др. Рассмотрим один из таких инструментов, который может применяться для обработки геопространственных данных – Google BigQuery GIS.

В хранилище данных, таком как BigQuery, информация о местоположении является очень востребованной. Многие важные бизнес-решения связаны с данными о местоположении. Например, можно записывать широту и долготу транспортных средств доставки или посылок с течением времени или же записывать транзакции клиентов и присоединять данные к другой таблице с данными о местоположении магазина. Геопространственная аналитика позволяет анализировать и визуализировать геопространственные данные в BigQuery с помощью географических типов данных и стандартных функций Google SQL geography. BigQuery – это хранилище данных Google со встроенными инструментами сбора, хранения и анализа географических данных. Для обработки сложных данных и изучения больших наборов данных он использует обычные SQL-запросы, которые часто применяются для обработки и визуализации географических данных [2].

Добавление данных в BigQuery осуществляется пакетами, загружая их или передавая напрямую в потоковом режиме, чтобы получать информацию в режиме реального времени (рисунок 1).

BigQuery использует множество встроенных функций, таких как геопространственный анализ, машинное обучение и бизнес-аналитика др., для сбора, хранения, анализа и визуализации данных. Геопространственная аналитика выявляет исторические и текущие изменения путем сбора, отображения и обработки изображений и данных, относящихся к определенному местоположению. Данные для BigQuery могут быть получены из GPS, мобильных устройств, датчиков местоположения, социальных сетей или спутниковых снимков. Собранная информация помогает создавать визуализации данных в виде графиков, карт, статистических таблиц и картограмм. Эти отчеты помогают человеку понять расстояние, близость и смежность, которые не видны в больших наборах данных.

BigQuery GIS – один из немногих программных продуктов, поддерживающий геопространственный анализ больших данных. В нем присутствует просмотр пространственных данных, поддержка произвольных точек, широт, полигонов и другие форматы геопространственных данных. Эта бессерверная архитектура упрощает процесс анализа. Данный продукт помогает анализировать геопространственные данные в больших масштабах, не требуя больших вычислительных мощностей.

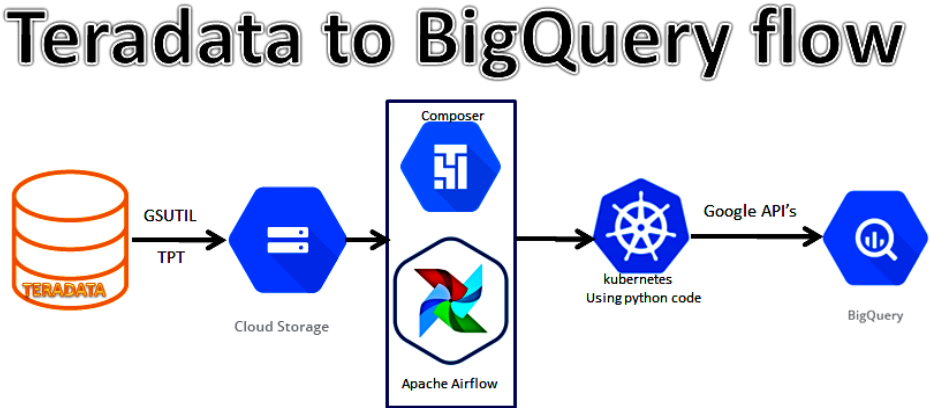

Рисунок 1. – Обработка больших данных в BigQuery

Функции BigQuery сгруппированы по следующим категориям:

1. Конструкторы – функции, которые создают новые географические значения на основе координат или существующих географических данных.
2. Синтаксические анализаторы – функции, которые создают географические данные из внешнего формата, такого как WKT и GeoJSON.
3. Форматирование – функции, которые экспортируют географические данные во внешний формат, такой как WKT.
4. Преобразования – функции, которые генерируют новую географию на основе входных данных.
5. Средства доступа – функции, которые обеспечивают доступ к свойствам географии без побочных эффектов.
6. Предикаты – функции, которые возвращают TRUE или FALSE для некоторой пространственной взаимосвязи между двумя географическими регионами или некоторым свойством географии. Эти функции обычно используются в предложениях фильтров.
7. Меры – функции, которые вычисляют измерения одной или нескольких географических областей.
8. Кластеризация – функции, которые выполняют кластеризацию по географическим регионам.

Ключевые особенности Google BigQuery.

– Полностью управляется Google: Google управляет инфраструктурой хранилища данных. Он поддерживает, обновляет, отслеживает и развертывает все данные или информацию.

– Простота реализации: не требуется никакого дополнительного программного обеспечения, развертывания кластера, виртуальных машин или инструментов с BigQuery. BigQuery – одно из экономичных бессерверных хранилищ данных. Для работы требуется загрузить или напрямую передать данные и запросы.

– Скорость: BigQuery может быстро обрабатывать массивы данных. Он может выполнять запросы на терабайты за секунды и петабайты за минуты.

BigQuery Geo Viz – один из веб-инструментов для визуализации геопространственных данных в BigQuery с использованием API Google Maps. Данный инструмент позволяет запускать SQL-запрос и отобразить результаты на интерактивной карте. Гибкие функции моделирования позволяют анализировать и изучать данные.

BigQuery Geo Viz не является полнофункциональным инструментом визуализации геопространственной аналитики. Geo Viz – это простой способ визуализации результатов запроса геопространственной аналитики на карте, по одному запросу за раз.

Еще одним из инструмента BigQuery является Google Data Studio. Google Data Studio – это бесплатный сервис для самостоятельного создания отчетов и визуализации данных от платформы Google, который подключается к BigQuery и сотням других источников данных. Сервис включает поддержку различных типов географических полей и картографических карт географических полигонов BigQuery. Визуализация на основе Карт Google позволяет отображать географические данные и работать с ними.

Кроме стандартного программного обеспечения и инструментов для реализации функций обработки Big Data имеется возможность создания собственных приложений на одном из языков программирования. Для этих целей в работе был выбран язык Phyton. На сегодняшнее время он считается универсальным языком программирования, который используется, в том числе для веб-разработки и создания специальных решений. Наибольшую популярность он приобрел в области обработки Big Data благодаря следующим преимуществам:

– низкий порог входа;

– множество готовых библиотек;

– наличие API в большинстве фреймворков.

Для примера в процессе работы был написан скрипт на Phyton, который позволяет извлекать о обрабатывать координаты постоянно действующих станций, используемых для изучения тектонической и геотермальной активности в штате Невада, а также для изучения глобальных закономерностей поверхностной нагрузки и проблем тектоники плит глобального масштаба. Данная информация доступна на сайте Невадской геодезической лаборатории [2]

Рассмотрим фрагмент данного скрипта и его основные функции.

```
import requests
from bs4 import BeautifulSoup
import csv

def get_html(url,params=None):
    r = requests.get(url,params=params)
    return r

def parse():
    html=get_html(URL)
    f.write (html.text)

f=open('e:/pars2.txt','w')
URL = 'http://geodesy.unr.edu/gps_timeseries/txyz/IGS14/00NA.txyz2'
parse()
f.close()
```

Опишем операторы и функции, которые были использованы в программном коде.

Вначале выполняем импорт трех библиотек: *requests, bs4, csv*

Requests - это модуль Python, который используется для отправки всех видов HTTP-запросов. Это простая в использовании библиотека с множеством функций, начиная от передачи параметров в URL-адресах до отправки пользовательских заголовков и проверки SSL.

BeautifulSoup4 (bs4) - это библиотека Python для извлечения данных из файлов HTML и XML. Для естественной навигации, поиска и изменения дерева HTML модуль BeautifulSoup4 по умолчанию использует встроенный в Python парсер html.

Библиотека *csv* используется для работы с форматом CSV (Comma Separated Values), который является одним из самых распространенных форматов импорта и экспорта электронных таблиц и баз данных. Функция *get_html* проверяет статус-код ресурса, т.е. узнаёт, работает ли сайт, адрес которого передается в нее как параметр. Инструкция *return* возвращает код запроса. Например, код 200 означает, что ресурс работает и можно с ним взаимодействовать. *Функция parse* получает информацию из веб-страницы и сохраняет ее в локальном файле, который создается оператором *f=open('d:/pars.txt','w').*

В заключении можно отметить, что большие данные уже меняют правила игры во многих сферах деятельности и, несомненно, их объем будет только увеличиваться, а технологии аналитики станут более совершенными. Большие данные – это одна из тех вещей, которые будут определять будущее человечества. В геодезии существует огромное количество данных, ручная обработка которых занимает много времени или вообще практически невозможна. Технология Big Data на порядок упрощает обработку этих данных.

ЛИТЕРАТУРА

1. Майер-Шенбергер, В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. – М. : Манн, Иванов и Фербер, 2014. – 240 с.
2. Лакшанан, В. Google BigQuery. Всё о хранилищах данных, аналитике и машинном обучении / В. Лакшанан, Д. Тайджани. – СПб. : Питер, 2021. – 469 с.
3. Аллен, Б. Д. Основы Python / Б. Д Аллен. – 2021. – С. 304.
4. Вайгенд, А. Big Data. Вся технология в одной книге/А.Вайгенд – «Эксмо» – 2017. - 384 с.
5. Вандер, Дж. Плас. Python для сложных задач наук о данных машинное обучение / Дж. Плас Вандер. – СПб. : Питер, 2018. – 576 с.
6. Что такое Google BigQuery и почему им стоит пользоваться [Электронный ресурс]. – Режим доступа: https://www.apix-drive.com/. – Дата доступа: 30.03.2022.