0619 ინფორმაციისა და კომუნიკაციის ტექნოლოგიები
INFORMATION AND COMMUNICATION TECHNOLOGIES (ICTS)

Apache Spark MLlib პლატფორმა მონაცემთა ტბებში მოთხოვნების დამუშავებისთვის

გიორგი მურადოვი
საქართველოს ტექნიკური უნივერსიტეტი
Georgian Technical University
E-mail: muradovi.giorgi22@gtu.ge

რეფერატი

დიდი მონაცემების ეპოქაში, მონაცემთა ტბები შეუცვლელი გახდა სტრუქტურირებული და არასტრუქტურირებული ტიპის დიდი მასშტაბის მონაცემების შესანახად. ტრადიციული მონაცემთა ბაზებისგან განსხვავებით, მონაცემთა ტბები ინახავს ნედლ მონაცემებს საკუთარ ფორმატში, რაც უზრუნველყოფს მოქნილობას მომავალი ანალიზისთვის. თუმცა, მონაცემთა მოცულობა და მრავალფეროვნება მონაცემთა ტბებში წარმოადგენს გამოწვევებს მოთხოვნის ეფექტური დამუშავებისთვის. ნაშრომში წარმოდგენილა მონაცემთა ტბების, როგორც ცენტრალიზებული საცავის როლი და მნიშვნელობა მასშტაბური მონაცემების შენახვისა და ანალიზისთვის. ასევე მოცემულია მონაცემთა ტბების სირთულეები დამუშავების თვალსაზრისით. აღნიშნულ პრობლემასთან გამკლავების მიზნით წარმოდგენილია Apache Spark ინსტრუმენტის როლი, როგორც მოქნილი ანალიტიკური მექანიზმი მონაცემთა ფართომასშტაბიანი დამუშავებისთვის. აღწერილია Apache Spark MLlib უპირატესობები მონაცემთა ტბებში მოთხოვნების დამუშავების პრობლემების გადასალახად. წარმოდგენილია პლატფორმები და სერვისები, რომლებიც იყენებენ Apache Spark MLlib-ს ინდუსტრიის აპლიკაციებს. ასევე განხილულია რეალური თანამედროვე აპლიკაციები და კვლევის მომავალი ტენდენციები და მიმართულებები.

საკვანძო სიტყვები: მოთხოვნის დამუშავება, მონაცემთა ტბები, Apache Spark, MLlib, Big Data Analytics, Machine Learning.

მონაცემთა ტბების სტრუქტურა და უპირატესობები

მონაცემთა ტბები ცენტრალიზებული საცავია, რომელიც ინახავს სტრუქტურირებულ, ნახევრად სტრუქტურირებულ და არასტრუქტურირებულ მონაცემებს ნედლ ფორმატში, სანამ ის საჭირო იქნება ანალიზისთვის ან სხვა ტიპის დამუშავებისთვის. ტრადიციული მონაცემთა ბაზებისგან განსხვავებით, რომლებსაც აქვთ მკაცრად განსაზღვრული სტრუქტურა, მონაცემთა ტბები საშუალებას იძლევა მონაცემთა შენახვა მოხდეს თავდაპირველ ფორმატში, რაც გაცილებით მეტ მოქნილობას სთავაზობს მომხმარებელს და შეიცავს მონაცემთა ტიპებისა და წყაროების მრავალფეროვნებას.

აღსანიშნავია მონაცემთა ტბების ძირითადი უპირატესობები [1]:

მასშტაბურობა: მონაცემთა ტბებს შეუძლიათ ჰორიზონტალურად მასშტაბირება, რათა უზრუნველყონ მზარდი მონაცემების მოცულობა.

მოქნილობა: ისინი მხარს უჭერენ მონაცემთა სხვადასხვა ფორმატსა და ტიპს, რაც საშუალებას აძლევს ორგანიზაციებს მიიღონ და შეინახონ მონაცემები წინასწარი სქემის განსაზღვრის გარეშე.

ხარჯების ეფექტურობა: ღრუბელზე დაფუძნებული შენახვის გადაწყვეტილებების და ღია წყაროს ტექნოლოგიების გამოყენება, როგორიცაა Apache Hadoop და Apache Spark, ამცირებს ინფრასტრუქტურის ხარჯებს მონაცემთა შენახვის ტრადიციულ მიდგომებთან შედარებით. მიუხედავად მათი სარგებლისა, მონაცემთა ტბები ხასიათდება გარკვეული სირთულეებით, მათ შორის აღსანიშნავია:

მონაცემთა მენეჯმენტი: მეტამონაცემების მართვა და მონაცემთა ხარისხისა და მართვის უზრუნველყოფა კრიტიკულია.

მოთხოვნის დამუშავება: ჰეტეროგენული მონაცემების დიდი მოცულობის ეფექტურად მოძიება მოითხოვს დამუშავების მყარ ჩარჩოებს და ოპტიმიზებულ ტექნიკას.

სირთულე: მონაცემთა მრავალფეროვანი ფორმატის მართვა, სქემის ევოლუცია და მონაცემთა რეალურ დროში დამუშავება მოითხოვს დახვეწილ გადაწყვეტილებებს.

Apache Spark MLlib, როგორც მძლავრი ინსტრუმენტი მონაცემთა დამუშავებისთვის

Apache Spark-მა მოიპოვა პოპულარობა, როგორც ერთიანი ანალიტიკური მექანიზმი მონაცემთა ფართომასშტაბიანი დამუშავებისთვის. ის უზრუნველყოფს მეხსიერების გამოთვლით ჩარჩოს, რომელიც მხარს უჭერს სხვადასხვა დატვირთვას, როგორიცაა სერიული დამუშავება, რეალურ დროში ნაკადის დამუშავება, ინტერაქტიული მოთხოვნები და მანქანური სწავლება. Spark-ის უნარი გაანაწილოს მონაცემთა დამუშავების ამოცანები კვანძების კლასტერში აძლიერებს შესრულებას და მასშტაბურობას [2,3].

MLlib არის Apache Spark-ის მასშტაბირებადი მანქანური სწავლების ბიბლიოთეკა, რომელიც შექმნილია მანქანური სწავლების მოდელების მასშტაბის გამარტივებისა და დაჩქარების მიზნით. ის მომხმარებელს სთავაზობს ალგორითმებისა და უტილიტების ფართო სპექტრს კლასიფიკაციის, რეგრესიის, კლასტერიზაციის, ერთობლივი ფილტრაციისა და განზომილების შემცირებისთვის. MLlib იყენებს Spark-ის განაწილებულ გამოთვლის შესაძლებლობებს, რაც მას შესაფერს ხდის მონაცემთა დიდი ნაკრების დასამუშავებლად და რთული ანალიტიკური ამოცანების ეფექტურად შესასრულებლად. სხვა პლატფორმებთან შედარებით დიდ მონაცემთა დამუშავების პლატფორმების ლანდშაფტში, Apache Spark MLlib გამოირჩევა რიგი უპირატესობებით, როგორიცაა:

შესრულება: Spark-ის მეხსიერებაში გამოთვლითი და ეფექტური მოთხოვნის დამუშავების შესაძლებლობები მნიშვნელოვნად ამცირებს მონაცემთა დამუშავების დროს დისკზე დაფუძნებულ სისტემებთან შედარებით.

გამოყენების სიმარტივე: Spark-ისა და MLlib-ის მიერ მოწოდებული მაღალი დონის API-ები ამარტივებს მონაცემთა რთული დამუშავებისა და მანქანური სწავლების სამუშაო პროცესების განხორციელებას.

მასშტაბურობა: Spark-ის უნარი გაანაწილოს გამოთვლები კლასტერზე, საშუალებას აძლევს მას ეფექტურად გაატაროს რამდენიმე პეტაბაიტი მონაცემები.

მონაცემთა ტბებში მოთხოვნების დამუშავების გამოწვევები

მონაცემთა ტბებში მოთხოვნის ეფექტური დამუშავება მოიცავს რამდენიმე გამოწვევის გადალახვას:

მოცულობა და მრავალფეროვნება: მონაცემთა ტბები ინახავს მონაცემთა მრავალფეროვან კომპლექტს, რომელთაც აქვთ, როგორც სტრუქტურირებული, ასევე არასტრუქტურირებული ფორმატი, რაც მოითხოვს მოთხოვნების დამუშავების მრავალმხრივ შესაძლებლობებს.

რეალურ დროში მოთხოვნები: ორგანიზაციები სულ უფრო მეტად ითხოვენ რეალურ დროში ანალიტიკას და დამუშავების შესაძლებლობებს თავიანთი მონაცემთა ტბებიდან, რაც საჭიროებს მოთხოვნის მინიმალური შეფერხებით შესრულებას.

მონაცემთა ხარისხი და მართვა: მონაცემთა ხარისხისა და მართვის უზრუნველყოფა მოთხოვნის დამუშავების სასიცოცხლო ციკლის განმავლობაში გადამწყვეტია ზუსტი და სანდო ინფორმაციის მისაღებად.

მოთხოვნის ეფექტური დამუშავების პრობლემის გადასაჭრელად შეიძლება გამოყენებულ იქნას რამდენიმე ტექნიკა:

მონაცემთა ინდექსირება და დაყოფა: მონაცემების ინდექსირება და დაყოფა ოპტიმიზაციას უკეთებს მოთხოვნის შესრულებას დასკანირებული მონაცემების მოცულობის მინიმიზაციის გზით.

ოპტიმიზაციის სტრატეგიები: როგორიცაა სვეტოვანი შენახვის ფორმატები, რომელიც აუმჯობესებს მოთხოვნის შესრულების ეფექტურობას.

Spark SQL: Spark SQL იძლევა სტრუქტურირებული მონაცემების მოთხოვნის საშუალებას SQL-ის გამოყენებით, შეუფერხებლად ინტეგრირდება Spark's DataFrame API-სთან მონაცემთა მანიპულაციისა და აგრეგაციისთვის [4].

პლატფორმები და სერვისები, რომლებიც იყენებენ Apache Spark MLlib-ს ინდუსტრიის აპლიკაციებს.

Apache Spark MLlib პოულობს ფართო გამოყენებას სხვადასხვა ინდუსტრიაში:

ელექტრონული კომერცია: საცალო ვაჭრობა იყენებს MLlib-ს პერსონალიზებული რეკომენდაციებისა და მომხმარებლების სეგმენტაციისთვის, შეძენის ისტორიისა და დათვალიერების ქცევის საფუძველზე.

ჯანმრთელობა: ჯანდაცვის პროვაიდერები იყენებენ MLlib-ს პროგნოზირებადი ანალიტიკისთვის, დაავადების პროგნოზირებისთვის და პაციენტის რისკის სტრატიფიკაციისთვის, ჯანმრთელობის ელექტრონული ჩანაწერების (EHR) მონაცემების გამოყენებით.

ფინანსები: ფინანსური ინსტიტუტები იყენებენ MLlib-ს თაღლითობის გამოვლენისთვის, საკრედიტო რისკის შეფასებისთვის და ალგორითმული ვაჭრობის სტრატეგიებისთვის, რომელიც ეფუძნება რეალურ დროში ბაზრის მონაცემებს.

განხილულიდან გამომდინარე ჩანს თუ რამდენად ეფექტურია Apache Spark Mllib ინსტრუმენტის გამოყენება სხვადასხვა მიმართულებით მასშტაბური ამოცანების გადასაწყვეტად. ამავდროულად, Apache Spark MLlib კონკურენციას უწევს სხვა დიდი მონაცემთა პლატფორმებს და სერვისებს, როგორიცაა:

Hadoop: მიუხედავად იმისა, რომ Hadoop უზრუნველყოფს განაწილებული შენახვისა და დამუშავების შესაძლებლობებს, Spark-ის მეხსიერებაში გამოთვლითი მოდელი გვთავაზობს უმაღლეს შესრულებას განმტკიცებებით მანქანური სწავლებისა და რეალურ დროში ანალიტიკისთვის.

Cloud Services (AWS EMR, Azure Databricks): მართული სერვისები, როგორიცაა AWS EMR და Azure Databricks, გვთავაზობს ინტეგრაციას Apache Spark-თან, უზრუნველყოფს მასშტაბირებულ ინფრასტრუქტურას და დამატებით მართულ სერვისებს მონაცემთა ტბის განლაგებისა და მანქანური სწავლების მოდელის ტრენინგისთვის.

რეალური თანამედროვე აპლიკაციები

რამდენიმე ორგანიზაციამ წარმატებით გამოიყენა Apache Spark MLlib მოთხოვნის დამუშავებისა და მანქანური სწავლისთვის მონაცემთა ტბებში:

Netflix: იყენებს Spark MLlib-ს კონტენტის რეკომენდაციის ალგორითმებისთვის, მომხმარებლის ჩართულობისა და შინაარსის პერსონალიზაციის გასაუმჯობესებლად.

Uber: იყენებს Spark MLlib-ს რეალურ დროში მონაცემთა ანალიტიკისთვის და პროგნოზირებადი მოდელირებისთვის, მგზავრობის სერვისების ოპტიმიზაციისთვის და მძღოლების განაწილებისთვის.

Airbnb: ახორციელებს MLlib-ს ფასების ოპტიმიზაციისა და მოთხოვნის პროგნოზირებისთვის, დაჯავშნის ისტორიული მონაცემებისა და ბაზრის ტენდენციების საფუძველზე.

კვლევები მიუთითებს შესრულების მნიშვნელოვან გაუმჯობესებასა და მასშტაბურობის უპირატესობებზე Apache Spark MLlib-ის გამოყენებისას მოთხოვნის დამუშავებისა და მანქანური სწავლების ამოცანებისთვის დამუშავების ტრადიციულ სისტემებთან შედარებით. ფართომასშტაბიანი მონაცემთა ნაკრების ეფექტურად დამუშავებისა და რეალურ დროში ქმედითი შეხედულებების გამომუშავების შესაძლებლობა აძლიერებს გადაწყვეტილების მიღების შესაძლებლობებს ინდუსტრიებში [5,6].

მომავალი ტენდენციები და კვლევის მიმართულებები

მომავალი მიღწევები Apache Spark MLlib-ში და მონაცემთა ტბის მოთხოვნის დამუშავებაში მოიცავს:

ხელოვნური ინტელექტისა და ღრმა სწავლების ინტეგრაციას: მანქანური სწავლების მოწინავე ტექნიკისა და ღრმა სწავლების მოდელების ჩართვა Spark MLlib-ში რთული ნიმუშის ამოცნობისა და პროგნოზირებადი ანალიტიკისთვის.

რეალურ დროში დამუშავება: Spark-ის შესაძლებლობების გაძლიერება მონაცემთა რეალურ დროში დამუშავებისა და ნაკადის ანალიტიკისთვის, რათა დააკმაყოფილოს მზარდი მოთხოვნა დაბალ დაყოვნებებთან მიმართებით. თუმცა ამასთან ერთად ძალაში რჩება გაფართოვების შესაძლებლობის არსებობა, ასევე მონაცემთა კონფიდენციალობისა და უსაფრთხოების საკითხების გამკაცრება.

ამდენად, Apache Spark MLlib თამაშობს გადამწყვეტ როლს მონაცემთა ტბებში მოთხოვნის ეფექტური დამუშავებისა და გაფართოებული ანალიტიკის განხორციელებაში. მისი ინტეგრაცია Spark-ის განაწილებულ გამოთვლით ჩარჩოსთან აძლევს ორგანიზაციებს საშუალებას გამოიყენონ ფართომასშტაბიანი მონაცემები ყოველდღიურ საქმიანობაში და ბიზნეს ანალიტიკისთვის. მონაცემთა მოცულობისა და სირთულის ზრდასთან ერთად, Apache Spark MLlib რჩება ინოვაციების წინა პლანზე, რაც განაპირობებს წინსვლას დიდი მონაცემების ანალიტიკასა და მანქანური სწავლებაში.

ლიტერატურა

1. Zaharia, M., et al. (2016). "Apache Spark: A Unified Analytics Engine for Big Data Processing." ACM SIGMOD entry.
2. Meng, X., et al. (2016). "MLlib: Machine Learning in Apache Spark." Journal of Machine Learning Research.
3. Chen, Y., et al. (2015). "Comparison of Big Data Platforms: Apache Spark vs. Hadoop." IEEE Transactions on Big Data.
4. AWS EMR documentation: https://aws.amazon.com/emr/
5. Azure Databricks documentation: https://azure.microsoft.com/en-us/services/databricks/
6. Li, H., et al. (2017). "Use cases of Apache Spark in industrial applications." IEEE International Conference on Big Data.

Apache Spark MLlib platform capabilities for query processing in data lakes

Giorgi Muradovi

Abstract

In the era of big data, data lakes have become indispensable for storing large-scale data of both structured and unstructured types. Unlike traditional databases, data lakes store raw data in their own format, providing flexibility for future analysis. However, the volume and variety of data in data lakes present challenges for efficient query processing. The paper presents the role and importance of data lakes as a centralized repository for large-scale data storage and analysis. The difficulties of data lakes in terms of processing are also presented. In order to deal with this problem, the role of the Apache Spark tool as a flexible analytical mechanism for large-scale data processing is presented. Describes the advantages of Apache Spark MLlib for overcoming query processing problems in data lakes. Platforms and services that use Apache Spark MLlib for industry applications are presented. Real modern applications and future research trends and directions are also discussed.

Keywords: Query Processing, Data Lakes, Apache Spark, MLlib, Big Data Analytics, Machine Learning.