ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი

ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი

Georgian Cartoon Films Portal - Big Data Analysis, Processing and Visualization

(ქართული მულტფილმების პორტალი დიდი მონაცემების ანალიზი, დამუშავება და ვიზუალიზაცია)

გიორგი კობაიძე

სამაგისტრო პროგრამა - ინფორმაციული ტექნოლოგიები

ნაშრომი შესრულებულია ინფორმაციული ტექნოლოგიების მაგისტრის აკადემიური ხარისხის მოსაპოვებლად

ხელმძღვანელი: მანანა ხაჩიძე

თბილისი

2019

რეზიუმე

პროექტი „რეანიმაცია“ არის ამ ნაშრომის ძირითადი ამოსავალი წერტილი. პროექტის მიზანია შეიქმნას პორტალი, სადაც იქნება ყველა არსებული ქართული მულტფილმი, მომხმარებლებს კი შესაძლებლობა ექნებათ შევიდნენ, ნახონ სასურველი მულტფილმი, გაეცნონ ინფორმაციას ზოგადად კინემატოგრაფიის შესახებ, დაწერონ შეფასება თითოეულ მულტფილმზე და ა.შ. გარდა ამისა თითოეულ მომხმარებელს შეეძლება ხელი შეუწყოს დაზიანებული მულტფილმების აღდგენას დონაციით. მოცემულ პროექტში მთავარ სირთულეს წარმოადგენს დიდი რაოდენობის ვიდეო მონაცემების შენახვა და მათი სტრიმინგი. სწორედ ამ თემაზეა მოცემული ნაშრომი, სადაც განვიხილავთ დიდ მონაცემებს სიღრმისეულად.

გასაღები სიტყვები

Big data, Streaming, Video, Technologies, Visualization, Classification, Clustering

(დიდი მონაცემები, სტრიმინგი, ვიდეო, ტექნოლოგია, ვიზუალიზაცია, კლასიფიკაცია, კლასტერიზაცია)

ნაშრომი სისტემატიზირებულია შემდეგნაირად:

სექცია 1-ში განხილულია დიდი მონაცემების (ვიდეო ან აუდიო ფაილების ჩვენს შემთხვევაში) სტრიმინგის მეთოდები, სექცია 2 გაგვაცნობს დიდი მონაცემების საჭიროებას, აპლიკაციებს, უპირატესობებს და მახასიათებლებს. დიდი მონაცემების დამუშავებისთვის გამოყენებული ხელსაწყოები და ტექნოლოგიები განხილულია სექცია 3-ში, სექცია 4-ში განხილულია სხვადასხვა კვლევის პრობლემა. სექცია 5 მოიცავს ვიზუალიზაციის ხელსაწყოებს, ბოლოს სექცია 6-ში მოყვანილია შეჯამება და განხილულია უახლესი ტენდენციები.

შესავალი:

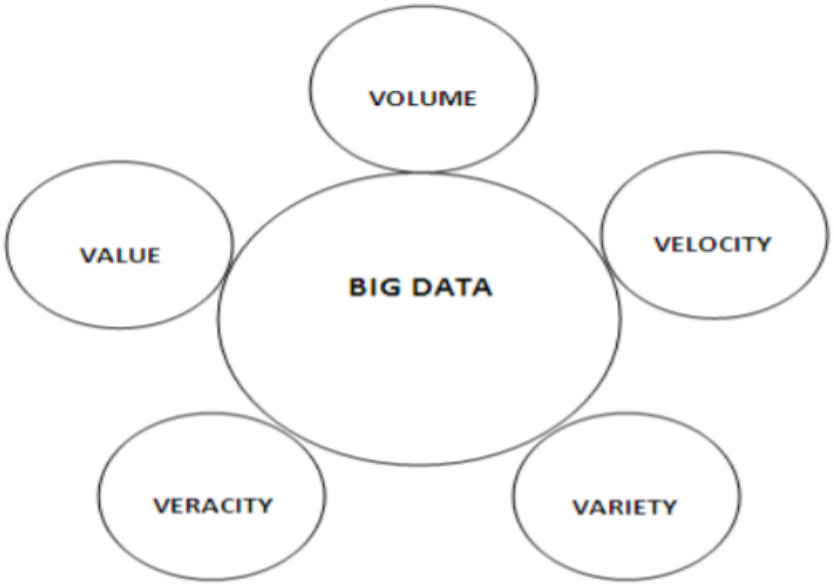
დიდი მონაცემები (Big data) არის მნიშვნელოვანი ტერმინი, რომელიც აღწერს მონაცემების განვითარებას და ხელმისაწვდომობას 3 ფორმატში, როგორიცაა სტრუქტურიზებული, არასტრუქტურიზებული და ნახევრად სტრუქტურიზებული. სტრუქტურიზებული მონაცემები ლოკალიზებულია ფიქსირებულ ველში, ჩანაწერში ან ფაილში და წარმოდგენილია რელაციურ მონაცემთა ბაზებში და ცხრილებში მაშინ, როცა არასტრუქტურიზებული მონაცემების ფაილები მოიცავს ტექსტს და მულტიმედიურ მონაცემებს. დიდი მონაცემების კონცეფტის მთავარი დანიშნულება არის ექსტრემალური მოცულობის მონაცემთა ნაკრებების აღწერა, როგორც სტრუქტურიზებულის, ასევე არასტრუქტურიზებულის. ეს ასევე განისაზღვრება სამი “V” განზომილებით: მოცულობა (Volume), სიჩქარე (Velocity), და მრავალფეროვნება (Variety), ემატება

კიდევ ორი “V”, როგორიცაა ღირებულება (Value) და სიზუსტე (Veracity). მოცულობა (Volume) აღნიშნავს მონაცემების ზომას, სიჩქარე (Velocity) - მონაცემების დამუშავების სიჩქარეს, მრავალფეროვნება (Variety) აღიწერება როგორც განსხვავებული მონაცემთა ტიპები, ღირებულება (Value) გამოხატავს ბიზნეს ღირებულებას და სიზუსტე (Veracity) აღნიშნავს მონაცემების ხარისხს და მათი აღქმის შესაძლებლობას. დიდი მონაცემები გახდა მეცნიერული კვლევებისთვის უნიკალური და პოპულარული განხრა კომპიუტერულ მეცნიერებაში. ამ განხრით მრავალი გადაუჭრელი ამოცანაა და მრავალი კარგი ამოხსნა არის შემოთავაზებული მკვლევარების მიერ. აუცილებელია ბევრი ტექნიკის და ალგორითმის განვითარება დიდი მონაცემებისთვის, რათა მივიღოთ ბევრად უკეთესი და ოპტიმალური გადაწყვეტები. ამ ნაშრომში განხილულია დიდი მონაცემების შესწავლის მეთოდები, საბაზისო კონცეფტები, ისტორია, აპლიკაციები, ტექნიკა, კვლევის პრობლემები და ხელსაწყოები.

გარდა დიდი მონაცემებისა, ასევე მნიშვნელოვანია ის ტექნიკა, თუ როგორ ხდება ქსელში ერთმანეთთან დაკავშირებულ კვანძებს შორის ინფორმაციის გაცვლა. როგორ ხდება ინფორმაციის წყაროდან მომხმარებლის მოწყობილობამდე მედია ნაკადის მისვლა მინიმალური დაყოვნებით და მაქსიმალური სიზუსტით.

დიდი მონაცემები ასოცირდება მონაცემების დიდ ნაკრებებთან და ზომებთან, რომელთა შენახვა და დამუშავება აღემატება სტანდარტული მონაცემთა ბაზების შესაძლებლობებს. დიდი მონაცემების ანალიზი არის ძალიან მნიშვნელოვანი ანალიტიკოსებისთვის, მკვლევარებისთვის და ბიზნესში მომუშავე ადამიანებისთვის რომ მიიღონ უკეთესი გადაწყვეტილებები სტატისტიკაზე დაყრდნობით. სურათი 1-ზე გამოსახულია დიდი მონაცემების სტრუქტურა, რომელიც შეიცავს ხუთ განზომილებას: მოცულობა, სიჩქარე, მრავალფეროვნება, ღირებულება და სიზუსტე. მოცულობა აღნიშნავს მონაცემთა ზომას, რომელიც გვიჩვენებს როგორ დავამუშაოთ დიდი მოცულობის მონაცემები დიდ მასშტაბიანი და მრავალ განზომილებიანი მონაცემთა ბაზებით, ასევე მისი დამუშავების საჭიროებას. სიჩქარე განსაზღვრავს მონაცემთა უწყვეტ ნაკადს. დიდ მონაცემებს აქვს გაუმჯობესებული სიჩქარე, კავშირი და გამოთვლის სისწრაფე.

მრავალფეროვნება განსაზღვრავს ინფორმაციის ხარისხს სხვადასხვა ადგილიდან, რაც ნიშნავს იმის განსაზღვრას თუ როგორ უნდა მივიღოთ სხვადასხვა ტიპის მონაცემები, მაგალითად მონაცემთა წყარო შეიცავს არა მხოლოდ სტრუქტურიზებულ ტრადიციულ რელაციურ მონაცემებს, არამედ ის ასევე შეიცავს მოჩვენებით სტრუქტურიზებულ, ნახევრად სტრუქტურიზებულ და არასტრუქტურიზებულ მონაცემებს, როგორიცაა ტექსტი, სენსორების მონაცემები, აუდიო, ვიდეო, გრაფი და სხვა მრავალი ტიპი. ღირებულება არის აუცილებელი რომ მივიღოთ ეკონომიკური ღირებულება სხვადასხვა მონაცემებით, რომლებიც აშკარად განსხვავდება ერთმანეთისგან. აქ მთავარი გამოწვევა არის ის რომ განვსაზღვროთ, რომელი არის ღირებული და შესაბამისი ტრანსფორმაციის ტექნიკა იმისთვის რომ შევასრულოთ მონაცემთა ანალიზი.



სურათი 1: დიდი მონაცემების სტრუქტურა

დიდი მონაცემების ერთ-ერთი ტიპიური მახასიათებელი არის სტრუქტურიზებული მონაცემების, ნახევრად სტრუქტურიზებული მონაცემების და არასტრუქტურიზებული მონაცემების ინტეგრაცია. დიდი მონაცემები იზომება სიჩქარით, ხარისხით, უსაფრთხოებით, მოხერხებულობით და სტაბილურობით. სხვა მნიშვნელოვანი უპირატესობა დიდი მონაცემების არის მონაცემთა ანალიზი. დიდი მონაცემების ანალიტიკა ნიშნავს შეგროვების, ორგანიზაციის, და ანალიზის პროცესს, რათა აღმოვაჩინოთ გარკვეული სქემა ან სხვა მნიშვნელოვანი ინფორმაცია. ცხრილი 1 გვიჩვენებს სხვადასხვა ტიპის და ზომის მონაცემების შესწავლის ტექნიკას, ხელსაწყოებს და მეთოდებს.

Data Types	Data Sizes	characteristics	Tools	Analytical methods	Examples
Small Data	Mega bytes	Hundred – thousand records	Personal computers, excel, R	Simple statistics	Sales records, customer Database for small companies
Large data	Giga bytes, Tera bytes	Millions of records - structured data	RDBMS, Data warehouses	Advance statistics, Data mining, business intelligence	customer Database for big companies
Big data	Giga bytes , Peta bytes	Over millions of records – distributed and unstructured	Cloud ,Data centre, NoSQL, Hadoop	Map reduce , Distributed file systems	Customer interaction-social network, mobile, multimedia

ცხრილი 1: სხვადასხვა ტიპის მონაცემების შესწავლის ტექნიკა

1. დიდი მონაცემების სტრიმინგის მეთოდები:

რას ნიშნავს დიდი მონაცემების სტრიმინგი?

დიდი მონაცემების სტრიმინგი არის პროცესი, რითაც დიდი მონაცემები სწრაფად პროცესირდება იმისათვის რომ დამუშავების შედეგი რეალურ დროში შესრულდეს და მომხმარებელმა ნახოს შედეგი. დიდი მონაცემების სტრიმინგი არის იდეალურად სიჩქარეზე ორიენტირებული მიდგომა, რომელიც გულისხმობს მონაცემთა განგრძობითი ნაკადის დამუშავებას.

ამ მიდგომის მთავარ სირთულეს ის წარმოადგენს, რომ მონაცემთა ნაკადი შეიძლება იყოს ძალიან სწრაფი და ამავდროულად ძალიან მძიმე, თავიდან ხდება მონაცემის მოთხოვნა, მდგომარეობის აღმოჩენა (condition detection) დროის მცირე მონაკვეთში (მილიწამებიდან წუთებამდე), ეს არ ეხება მხოლოდ ვიდეო ან აუდიო სტრიმს, მარტივ მაგალითად შეგვიძლია მოვიყვანოთ სისტემა, რომელიც რეალურ დროში ამუშავებს ინფორმაციას ტემპერატურას და როცა ტემპერატურა მიაღწევს x გრადუსს უნდა გაგვაფრთხილოს.

ამ მიდგომას სხვა ალტერნატიული სახელწოდებებიც გააჩნია: ანალიტიკა რეალურ დროში (real-time analytics), სტრიმინგის ანალიტიკა (streaming analytics), კომპლექსური ივენთების პროცესინგი (complex event processing), ივენთების პროცესინგი (event processing) და ა.შ. ეს ტექნოლოგია პოპულარიზებულია Apache Storm-ის მიერ როგორც „Hadoop-ის მსგავსი ტექნოლოგია, რომელიც შედეგს იძლევა უფრო სწრაფად“ (“Technology like Hadoop but can give you results faster”). ამის შემდეგ ის მიიჩნევა როგორც დიდ მონაცემთა ტექნოლოგია.

რისთვის არის ნაკადის პროცესინგი საჭირო?

მიზეზი 1:

ზოგჯერ მონაცემები წარმოდგენილია უწყვეტი და დაუსრულებელი ნაკადის სახით, რაც აბსოლუტურად ბუნებრივია დღევანდელი მოცემულობით. სტანდარტული, ნაწილობითი პროცესინგის შემთხვევაში სისტემას უწევს შეინახოს მონაცემთა პარტია, დროებით ან მუდმივ მეხსიერებაში, შეაჩეროს ნაკადი, დაამუშაოს, გააგრძელოს ნაკადი და გადავიდეს შემდეგ პარტიაზე, ამ ყველაფერს ემატება ისიც, რომ მონაცემთა პარტიების გაერთიანება შეიძლება საკმაოდ კომპლექსური პროცედურა აღმოჩნდეს, რაც კიდევ უფრო ართულებს საერთო სურათს. სტრიმინგის საშუალებით შეგვიძლია ყველა ამ სირთულეს თავი ავარიდოთ, უფრო კონკრეტულად, აღმოვაჩინოთ მონაცემთა შაბლონები, გამოვიკვლიოთ შედეგები და ასევე შევხედოთ მონაცემებს სხვადასხვა პარალელური ნაკადებიდან ერთდროულად.

ნაკადის პროცესინგი ბუნებრივად თავსდება რეალურ დროზე დამოკიდებულ მონაცემებზე, რადგან, როგორც უკვე ვახსენეთ, ამ ტექნიკას აქვს შაბლონების აღმოჩენის უნარი, მაგალითად თუ ვცდილობთ დავადგინოთ ვებ სესიის ხანგრძლივობა დაუსრულებელ ნაკადში, ნაწილობითი დამუშავებით ამის გაკეთება ძალიან რთულია, რადგან ზოგიერთი სესია შეიძლება 2 ან მეტ მონაცემთა პარტიაში მოხვდეს, ხოლო ნაკადის პროცესინგის ტექნიკას ამის მოგვარება ძალიან მარტივად შეუძლია.

მიზეზი 2:

სტანდარტული დამუშავების მეთოდი მონაცემებს ტვირთავს მთლიანად და მხოლოდ ამის შემდეგ იწყებს მის დამუშავებას, ეს შეიძლება იყოს საკმაოდ ხანგრძლივი პროცესი, მიუხედავად იმისა, რომ თანამედროვე მსოფლიოში ყოველდღიურად უმჯობესდება პროცესორების სიმძლავრე და სისწრაფე. ფაქტობრივად, ეს ბუნებრივიცაა, რადგან რაც უფრო მძლავრი პროცესორები არსებობს მით უფრო დიდი ზომის და რაოდენობის მონაცემები იტვირთება გლობალურ ქსელში. სტრიმინგის საშუალებით კი შესაძლებელია მონაცემების ნაკადის დამუშავება, რაც არ საჭიროებს მონაცემთა ერთდროულად ჩატვირთვას, ამის გამო იზოგება პროცესორის სიმძლავრეც და პროცესინგის დროც, მაგრამ გასათვალისწინებელია ის ფაქტორიც რომ ნაკადის პროცესინგის მეთოდი არ უზრუნველყოფს აბსოლუტურად ზუსტ შედეგებს, რის გამოც ეს მიდგომა კარგია იმ შემთხვევისთვის, როცა მიახლოებითი შედეგი არის საკმარისი.

მიზეზი 3.

ზოგჯერ მონაცემები არის ძალიან დიდი და ვერ ხერხდება მისი შენახვა. ნაკადის პროცესინგი კი საშუალებას გვაძლევს დავამუშაოთ მთლიანი მონაცემები, მაგრამ შევინახოთ მხოლოდ კრიტიკულად აუცილებელი ნაწილები.

მიზეზი 4.

დღესდღეობით საკმაოდ მრავალი ტიპის მონაცემი არსებობს ისეთი, რომლის ნაკადად დამუშავება არის შესაძლებელი. გარდა ვიდეო და აუდიო სტრიმებისა არსებობს მომხმარებლის ტრანზაქციები, აქტივობები, ვებ-გვერდის ვიზიტები. მათი რაოდენობა დღითიდღე იზრდება IoT (Internet of Things)-ის დახმარებით.

მედია სტრიმინგის გადაცემის დიზაინი და იმპლემენტაცია:

ინტერნეტის სწრაფი დეველოპმენტი და პოპულარიზაცია განაპირობებს მედია სტრიმინგის ბიზნესის სწრაფ და ინერციულ განვითარებას. სტრიმინგ მედია ფართოდ გამოიყენება მულტიმედიურ სიახლეებში, ონლაინ ლაივ გადაცემებში, რეკლამირებაში, ტელემედიცინაში, ინტერნეტ რადიოში, ონლაინ კონფერენციებში და სხვა უმნიშვნელოვანეს საკითხებში. მედია სტრიმინგის ტექნოლოგია არაა უბრალოდ ერთი კონკრეტული ტექნოლოგია, ეს არის ქსელის და ვიდეო/აუდიო ტექნოლოგიების ორგანული კომბინაცია.

როგორ ხდება მედია სტრიმინგის გადაცემა?

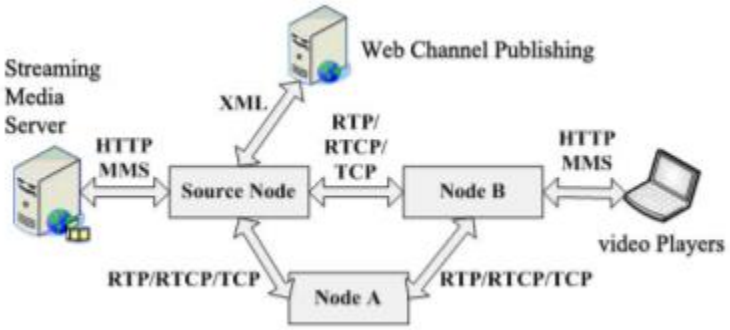
იმის გამო რომ IP ქსელის პაკეტების დაფორვარდების მეთოდი შექმნილია მხოლოდ მონაცემების წყვეტილად გადაცემისთვის, განგრძობითი მედია სტრიმების გადაცემა მისი გამოყენებით შეუძლებელია. იმისათვის რომ მოვახდინოთ მაღალი ხარისხის ვიდეო სტრიმების გადაცემა ქსელში ეფექტურად, უნდა გადავჭრათ რამდენიმე მნიშვნელოვანი ამოცანა. განვიხილოთ თითოეული მათგანი:

მედია სტრიმინგის გადაცემის სისტემის ზოგადი სტრუქტურა:

MixCast მედია სტრიმინგის გადაცემის სისტემაში, როცა ერთ-ერთი კვანძი გამოუშვებს გადაცემის სიგნალს, პირველად უკავშირდება იმ კვანძს, რომელსაც tracker-იპოვის თავისუფალს შემთხვევითი

შერჩევის გზით. როცა სტრიმინგ მედია სისტემა დაიწყებს გაშვებას, წარმოიქმნება ბუფერული დაყოვნება. ეფექტურ სტრატეგიად კი მიიჩნევა რომ ეს ბუფერული დაყოვნება უნდა იყოს რაც შეიძლება მცირე.

ზოგადი ტიპის მედია სტრიმის გადაცემის სისტემა იყენებს RTP პროტოკოლს, რომელიც თავის მხრივ იყენებს UDP პროტოკოლს ქვედა დონეზე, ასევე იყენებს RTCP პროტოკოლს რომ გააკონტროლოს პაკეტის ინფორმაცია, HTTP პროტოკოლს, რათა დაუკავშირდეს ვებ სერვერს. ვიდეო ფლეიერები მონაცემთა წყაროს უკავშირება HTTP და MMS პროტოკოლს, როგორც ნაჩვენებია ფიგურაზე



ვებ სერვერი იყენებს XML ტექნოლოგიას, კვანძი უკავშირდება ვებ სერვერს HTTP პროტოკოლით, იყენებს HTTP/MMS პროტოკოლს რომ დააკავშიროს მონაცემთა წყარო და ფლეიერი, კვანძი იღებს სტრიმინგ მედიის მონაცემებს ლოკალური ქეშიდან ან მედია წყაროდან და შემდეგ იძახებს ვიდეო ფლეიერს რათა დაიწყოს ვიდეოს გაშვება.

თითოეულ კვანძს შორის სიგნალი გადაეცემა TCP პროტოკოლით, რათა დავრწმუნდეთ, რომ ინფორმაცია მიიღება ზუსტად, დანაკლისის გარეშე. რეალური მედიის მონაცემები კი გადაიცემა RTP პროტოკოლით. გამგზავნ კვანძთან პაკეტი ენკაფსულირებულია RTP-ს მეშვეობით, დამატებული აქვს დროის შტამპი და მიმდევრობის ნომერი, შემდეგ კი გადაიცემა UDP დონეზე. მიღების კვანძზე პროცესდება პირიქით. RTCP საკონტროლო პაკეტებით გადაიცემა მონაცემები UDP/TCP და IP-ს გავლით.

ბუფერის დიზაინი და დანიშნულება:

სისტემაში მას შემდეგ რაც მონაცემთა წყაროს კვანძი მიიღებს ვიდეოს მონაცემებს, ის მონაცემთა ნაკადს დაყოფს ფრაგმენტებად, რომელთა ზომა ერთმანეთის ტოლია, მისი პასუხისმგებლობაა, რომ თითოეულ ფრაგმენტს მიანიჭოს სერიული ნომერი, შემდეგ ამატებს ციკლურ რიგის ტიპის ბუფერში. საერთო კვანძი იღებს მედია მონაცემების ფრაგმენტებს ქსელის გავლით, შემდეგ კი ამატებს მათ ციკლურ ბუფერში.

MixCast სისტემებში თითოეული კვანძი აპლიკაციის დონეზე არის მრავალი-მრავალთან კავშირში, რაც იმას ნიშნავს, რომ კვანძმა შეიძლება მიიღოს მონაცემები რამდენიმე კვანძისგან ერთდროულად და ასევე კვანძმა შეიძლება გააგზავნოს პაკეტები 1-ზე მეტ კვანძზე. ეს პროცესი გაკონტროლებულია თითოეულ კვანძზე.

2. დიდი მონაცემების საჭიროება

მონაცემების მასიური მოცულობის დამუშავებას არ ითვალისწინებს ტრადიციული მონაცემთა ბაზების სტრატეგიები და ხელსაწყოები და ეს ძირითადად კონცენტრირებულია სტრუქტურიზებულ მონაცემებზე სამუშაოდ. ქსელის შექმნასთან ერთად მონაცემები, რომლებიც შენახულია კომპიუტერის მეხსიერებაში საგრძნობლად გაიზარდა, რადგან გაიზარდა მოწყობილობების განვითარების ტემპიც. ინტერნეტის განვითარებასთან ერთად დრამატულად გაიზარდა ფიზიკურ მეხსიერებაში შესანახი ინფორმაციის რაოდენობაც. ეს ინფორმაცია არის არაერთი ტიპის და დანიშნულების და გამოიყენება სხვადასხვა სფეროებში. მრავალი ახალი ტექნიკა, მეთოდი და კონცეპტი იყო შემუშავებული და შემოთავაზებული მკვლევარების მიერ სტატიკური მონაცემთა ნაკრებების ანალიზისთვის. ჩვენს ორობით ერაში, როცა უკვე შექმნილია მობილური და უსადენო ტექნოლოგიები, ძალიან გავრცელებულია სხვადასხვა სოციალური ქსელი, სადაც ხალხი ერთმანეთს უზიარებს საკუთარ ინფორმაციას. ასეთ საიტებს მიეკუთვნება: Facebook, Twitter, Instagram და სხვა მრავალი. მონაცემთა ბაზებში, რომლებიც მსგავსი სოციალური ქსელების უკან დგას ინფორმაცია შემოდის უწყვეტი ნაკადით და ძალიან დიდი სიჩქარით და ბუნებრივია, ასეთი ინფორმაციის შენახვისთვის არ იქნება საკმარისი სტანდარტული კომპიუტერის მეხსიერება, რადგან ეს მონაცემები არის სწორედ იმ ტიპის მონაცემები, რომელსაც ჩვენ ამ მომენტში განვიხილავთ - დიდი მონაცემები. ამ სიტუაციამ ასევე შექმნა ამოცანა, თუ როგორ უნდა შევასრულოთ მონაცემთა ანალიზი დინამიურ მონაცემთა ნაკრებში, რადგან იმ დროისთვის არსებული სტანდარტული ალგორითმები არ შეესაბამება დიდი მონაცემების დამუშავებას.

ტერმინი “დიდი მონაცემები” გამოჩნდა 1998 წელს Silicon Graphics-ში (SGI) ჯონ მეშის მიერ. დიდი მონაცემების ზრდა იწვევს დამგროვებელი მოწყობილობების ტევადობის და დამუშავების სიმძლავრის ზრდას. ხშირად მონაცემების დიდი რაოდენობა (2.5 ქვანტილიონი) იქმნება სოციალურ ქსელში. დიდი მონაცემების ანალიზი გამოიყენება ამ დიდი მონაცემების გამოსაცდელად და დამალული სქემების და კორელაციების აღმოსაჩენად. ორი ტექნოლოგია, რომელიც გამოიყენება დიდი მონაცემების ანალიზისთვის არის NoSQL და Hadoop. NoSQL არის არარელაციური მონაცემთა ბაზების ტექნოლოგია, რისი მაგალითებიცაა HBase, Cassandra და mongoDB. Hadoop არის eco software პაკეტი რომელიც შეიცავს HDFS და MapReduce-ს. ხელსაწყოები როგორიცაა SAS, R, და Matlab, მხარს უჭერს გადამწყვეტ ანალიზს, მაგრამ არ არის შექმნილი დიდი მონაცემთა ნაკრებებისთვის და არც DBMS ან Map Reduce-ს არ შეუძლია მართოს მონაცემები, რომლებიც შემოდის დიდი სიხშირით.

დიდი მონაცემების აპლიკაციებს მიეკუთვნება ფართო მასშტაბიანი აპლიკაციები, რომლებიც მუშაობენ დიდ მონაცემთა ნაკრებებზე. არსებული პროგრამული უზრუნველყოფები, რომლებიც განკუთვნილია დიდი მონაცემებისთვის, როგორიცაა Apache Hadoop და Google-ს map reduce framework აგენერირებს დიდი რაოდენობით შუალედურ მონაცემებს. დიდი მონაცემები გამოიყენება მრავალ სფეროში, როგორიცაა წარმოება, ბიოინფორმატიკა, ჯანდაცვა, სოციალური ქსელი, ბიზნესი, მეცნიერება და ტექნოლოგია.

3. დიდი მონაცემების ტექნოლოგიები

სვეტებზე ორიენტირებული მონაცემთა ბაზები

სვეტებზე ორიენტირებული მონაცემთა ბაზები მონაცემებს ინახავს სვეტებში და არა სტრიქონებში, რაც გამოიყენება მასიური მონაცემების კომპრესირებისთვის და სწრაფი query-ებისთვის.

უსქემო მონაცემთა ბაზები

უსქემო მონაცემთა ბაზებს ასევე ეწოდება NoSQL მონაცემთა ბაზები. მონაცემთა ბაზა გვთავაზობს მექანიზმს, რომელიც განსხვავდება ცხრილური რელაციური ბაზებისაგან. არსებობს ორი ტიპის მონაცემთა ბაზები როგორიცაა დოკუმენტის პრინციპით შემნახველი და key-value პრინციპით შემნახველი, რომელიც ინახავს და მოაქვს მასიური რაოდენობის სტუქტურიზებულ, არასტრუქტურიზებულ და ნახევრად სტრუქტურიზებულ მონაცემებს.

Hadoop

Hadoop არის პოპულარული open source ხელსაწყო დიდ მონაცემებთან სამუშაოდ. ეს არის Java-ზე დაფუძნებული პროგრამული ფრეიმვორქი, რომელიც მხარს უჭერს დიდი მონაცემების ნაკრებს დისტრიბუციულ კომპიუტინგში. Hadoop კლასტერი იყენებს master/slave სტრუქტურას. დისტრიბუციული ფაილური სისტემა საშუალებას აძლევს სისტემას რომ გააგრძელოს ნორმალური ოპერირება. Hadoop-ს აქვს ორი ძირითადი ქვე პროექტი, სახელად Map Reduce და Hadoop Distributed File System (HDFS).

Map Reduce

ეს არის პროგრამული პარადიგმა, რომელიც საშუალებას გვაძლევს დიდი პროგრამული დავალება გავანაწილოთ ათასობით სერვერზე და სერვერულ კლასტერებზე. Map reduce იმპლემენტაცია შედგება ორი დავალებისგან როგორიცაა map task და reduce task. Map task-ში შემავალი მონაცემთა ნაკრები გადაკონვერტირებულია სხვადასხვა key/value წყვილში.

HDFS

Hadoop დისტრიბუციული ფაილური სისტემა არის ფაილური სისტემა, რომელიც იშლება ყველა კვანძზე Hadoop კლასტერში მონაცემთა შესანახად. ის აკავშირებს ყველა ფაილურ სისტემას ლოკალურ კვანძზე, რათა შექმნას დიდი ფაილური სისტემა. იმისათვის რომ გადალახოს კვანძში წარმოშობილი ხარვეზები HDFS აფართოებს უსაფრთხოების დონეს მონაცემების გამოსახვით რამდენიმე წყაროში.

Hive

Hive არის მონაცემების საწყობის ინფრასტრუქტურა, რომელიც აწყობილია Hadoop-ის ბაზაზე. მას აქვს სხვადასხვა შესანახი ტიპები, როგორიცაა ტექსტი, RC ფაილი, Hbase, ORC და სხვა. ჩაშენებული მომხმარებლის მიერ განსაზღვრული ფუნქციები. გამოიყენება თარიღებთან, სტრინგებთან, და სხვა მონაცემების მაინინგის ხელსაწყოებთან სამუშაოდ.

შესანახი ტექნოლოგიები

იმისათვის რომ შევინახოთ დიდი მოცულობის მონაცემები, ოპტიმალური და ეფექტური ტექნიკა არის აუცილებელი. შესანახი ტექნოლოგიების მთავარი კონცენტრირების არე არის მონაცემთა შეკუმშვა და მეხსიერების ვირტუალიზაცია.

Hbase

Hbase არის განშლადი დისტრიბუციული მონაცემთა ბაზა, რომელიც იყენებს Hadoop დისტრიბუციულ ფაილურ სისტემას შესანახად. ის მხარს უჭერს სვეტებზე ორიენტირებულ მონაცემთა ბაზას და მონაცემთა სტრუქტურას.

Chukwa

Chukwa ანალიტიკა აკვირდება დიდ დისტრიბუციულ სისტემას და ამატებს აუცილებელ სემანტიკას ლოგირების კოლექციისთვის და იყენებს end-to-end delivery მოდელს.

4. კვლევის პრობლემები დიდ მონაცემებში

დიდ მონაცემებს აქვს სამი ფუნდამენტალური პრობლემა, როგორიცაა მეხსიერების პრობლემა, მენეჯმენტის პრობლემა და დამუშავების პრობლემა. ეს პრობლემები გვაჩვენებს ძალიან დიდი რაოდენობით ტექნიკური კვლევის პრობლემებს. მონაცემები იქმნება ზოგადად ყველა ადგილას, მაგალითად სოციალურ ქსელში 12+ ტერაბაიტი ზომის ტვიტები იქმნება დღეში და საშუალოდ მათი რე-ტვიტები არის 144 თითოეულ ტვიტზე. შემდეგი პრობლემა არის მენეჯმენტის პრობლემა, რომელიც არის ძალიან რთულად გადასაჭრელი დიდი მონაცემების დომეინში. თუ მონაცემები არის განაწილებული გეოგრაფიულად მისი მართვა შეიძლება სხვადასხვა არსით. დამუშავების პრობლემა მდგომარეობს იმაში თუ როგორ უნდა დავამუშავოთ 1K ტერაბაიტი მონაცემები, რომლებიც მოითხოვს მთლიან დამუშავების კოლოსალურ და არარეალურ დროს, როგორიცაა მაგალითად 635 წელი. აქედან გამომდინარე, რამდენიმე ეგზაბიტი მონაცემის ეფექტური დამუშავებისთვის საჭიროა პარალელური დამუშავების მეთოდის განვითარება და ახალი ანალიტიკური ალგორითმების შემუშავება.

4.1. დიდი მონაცემების კლასიფიკაცია

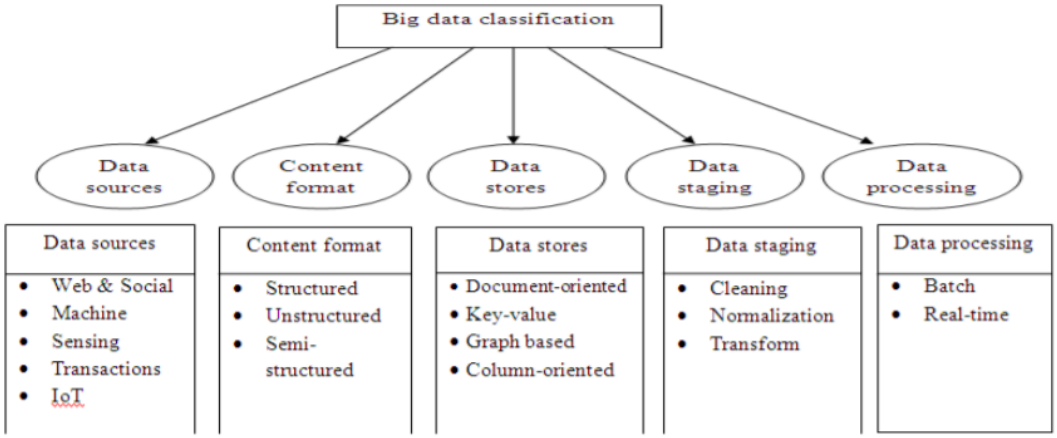
მონაცემთა კლასიფიკაცია არის მონაცემთა ორგანიზაციის პროცესი კატეგორიებად მათი რაც შეიძლება ეფექტურად და ოპტიმალურად გამოყენებისათვის. კარგად დაგეგმილი მონაცემთა კლასიფიკაციის სისტემა გვეხმარება რომ ადვილად მოვძებნოთ საჭირო მონაცემი. არსებობს სამი მთავარი ასპექტი მონაცემთა კლასიფიკაციისთვის, როგორიცაა მეთოდები, დომეინები და ვარიაციები. მეთოდები აღწერს ძირითად ტექნიკას, რომელიც გამოიყენება კლასიფიკაციისთვის, ამის მაგალითია ალბათობის მეთოდები, გადაწყვეტილების ხეები, წესზე დაფუძნებული მეთოდები, მაგალითზე დაფუძნებული მეთოდები, ვექტორული მანქანური მეთოდები და ნეირონული ქსელები.

დიდი მონაცემების ტიპების კლასიფიკაცია დაყოფილია სამ კატეგორიად: სოციალური ქსელი, ტრადიციული ბიზნეს სისტემები და Internet of Things. სოციალური ქსელი (ადამიანური ინფორმაცია) შეიცავს ინფორმაციას, რომელიც შედგება ადამიანური გამოცდილების ჩანაწერებისგან, ადრე ასეთი ჩანაწერები იყო წიგნებში და სამხატვრო ნამუშევრებში, ასევე აუდიო და ვიდეო ჩანაწერებში. ადამიანური ინფორმაცია ახლა არის თითქმის მთლიანად გაორობითებული და ჩაწერილი როგორც პერსონალურ კომპიუტერებში, ასევე სოციალურ ქსელებში.

Internet of Things (მანქანის მიერ დაგენერირებული მონაცემები) რაც მიიღება სენსორების და მანქანების რაოდენობის ფენომენალური ზრდით და ფიზიკური სამყაროს ივენთების და სიტუაციების მრავალი ჩანაწერით. ამ სენსორების გამომავალი ინფორმაცია არის მანქანურად დაგენერირებული მონაცემები, რომლებიც არის კარგად სტრუქტურიზებული. სენსორებიდან გამომავალი მონაცემები განსხვავდება ტიპების მიხედვით: ფიქსირებული სენსორები, სახლის ავტომატიზაციის მექანიზმები,

ამინდის ან ჰაერის დაბინძურების სენსორები, webcam სენსორები, სამეცნიერო სენსორები, ვიდეოები, მობილური სენსორები, სატელიტური სურათები და მონაცემები კომპიუტერის სისტემური ლოგებიდან და ვებ ლოგებიდან.

დიდი მონაცემების კლასიფიკაცია არის ხუთ ასპექტზე დაფუძნებული: მონაცემთა წყარო, შიგთავსის ფორმატი, მონაცემთა შემნახველები, მონაცემთა სთეიჯინგი და მონაცემთა დამუშავება. ეს წარმოდგენილია მეორე სურათზე. თითოეული კლასიფიკაცია მოითხოვს ალგორითმებს და ტექნიკას კლასიფიკაციის დავალების ოპტიმალურად შესასრულებლად დიდი მონაცემების დომეინში.

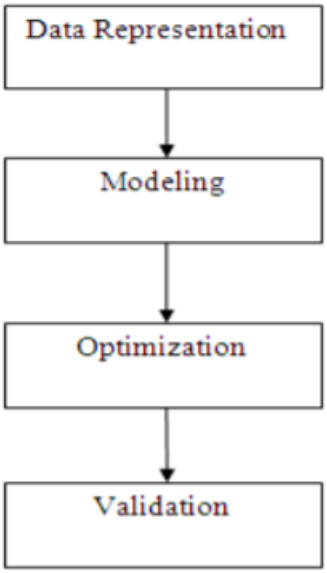


სურათი 2. დიდი მონაცემების კლასიფიკაცია

მონაცემთა წყარო არის მონაცემები შეგროვებული რამდენიმე ქვე-წყაროდან. მონაცემთა მნიშვნელოვან წყაროებს წარმოადგენს ვებ და სოციალური მედია, მანქანურად დაგენერირებული მონაცემები, სენსორების მონაცემები, ტრანზაქციების მონაცემები და Internet of Things. სოციალური მედია შეიცავს მოცულობით ინფორმაციას, რომელიც დაგენერირებულია URL-ის გამოყენებით რათა გავაზიაროთ ან გავცვალოთ ინფორმაცია ვირტუალურ კომუნიკაციებში და ქსელში. მაგალითად Facebook, Twitter ან ბლოგებში. მანქანურად გენერირებულ მონაცემებში ინფორმაცია არის ავტომატურად დაგენერირებული მოწყობილობებიდან და პროგრამული უზრუნველყოფიდან, მაგალითად კომპიუტერებიდან ან სამედიცინო მოწყობილობებიდან. ტრანზაქციის მონაცემები მოიცავს დროის განზომილებას რათა მოახდინოს მონაცემების ილუსტრაცია, მაგალითად ფინანსური და ბიზნეს მონაცემები.

4.2. კლასტერები დიდ მონაცემებში

ერთად არსებულ იდენტურ ელემენტების ჯგუფს ეწოდება კლასტერები. მონაცემთა კლასტერიზაცია არის ასევე ცნობილი, როგორც კლასტერების ანალიტიკა, რომელიც n ობიექტის კოლექციას გადააწყობს და დაალაგებს დანაყოფში ან იერარქიაში. კლასტერიზაციის მთავარი მიზანია მონაცემთა კლასიფიკაცია კლასტერებში, როგორიცაა ობიექტები, რომლებიც დაჯგუფებულია მსგავსობის და იდენტობის მიხედვით. ყველაზე გავრცელებული ალგორითმები კლასტერიზაციისთვის არის დანაყოფების, იერარქიული, ცხრილზე დაფუძნებული, სიმჭიდროვეზე დაფუძნებული და მოდელზე დაფუძნებული ალგორითმები. სურათი 3-ზე გამოსახულია მონაცემთა კლასტერიზაციის პროცესი.



სურათი 3. კლასტერიზაცია

4.3. დიდი მონაცემების ვიზუალიზაცია

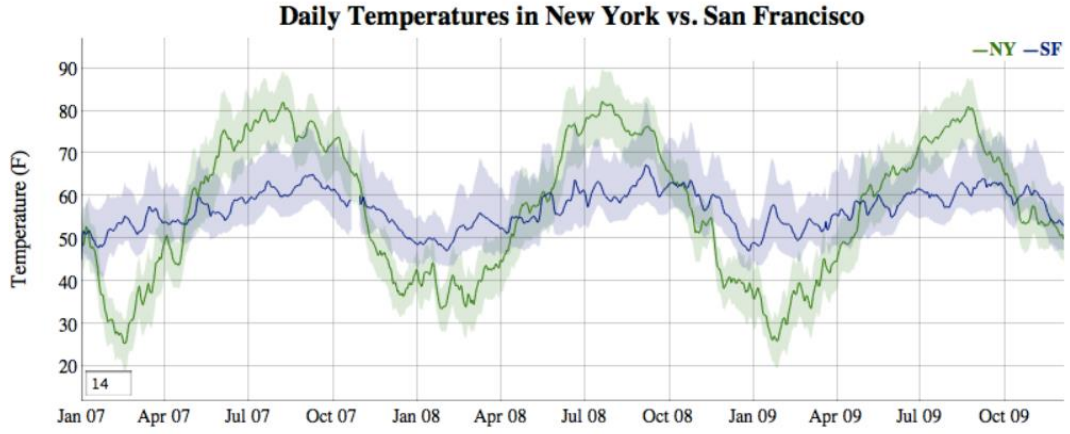
დიდი მონაცემების ვიზუალიზაცია არის რიცხვითი მონაცემის გამოსახვა 3D სურათებში. ეს არის გრაფიკული ფორმატით მონაცემის წარმოდგენა, რომელიც დამოკიდებულია ვიზუალურ კომპონენტებზე და გვეხმარება ადვილად და უფრო სწრაფად შევისწავლოთ სტატისტიკური მონაცემები. არსებობს მრავალი ხელსაწყო დიდი მონაცემების ვიზუალიზაციისთვის, როგორიცაა polymaps, nodebox, flot, processing, tangle, SAS visual anaytics, linkscape, leafelet, crossfilter, openlayer. ვიზუალიზაციის ტექნიკები არის კლასიფიცირებული სამი სხვადასხვა გზით, რომლებიც დაფუძნებულია დავალებაზე, მონაცემის სტრუქტურაზე ან განზომილებაზე. ვიზუალიზაცია შეიძლება კლასიფიცირდეს იმის მიხედვით მოწოდებული მონაცემი სივრცულია თუ არასივრცული, რადგან გამოვსახოთ შესაბამისი წესით ან 2D ან 3D. ვიზუალიზაციის კომპონენტები შეიძლება იყოს როგორც სტატიკური, ასევე დინამიური.

5. მონაცემთა ვიზუალიზაციის ხელსაწყოები

5.1. Dygraphs

Dygraph-ები არის სწრაფი, მრავალმხრივი, open source JavaScript ბიბლიოთეკა. მას აქვს პერსონალიზაციის დიდი შესაძლებლობა და შექმნილია მჭიდრო მონაცემთა ნაკრების ვიზუალურად

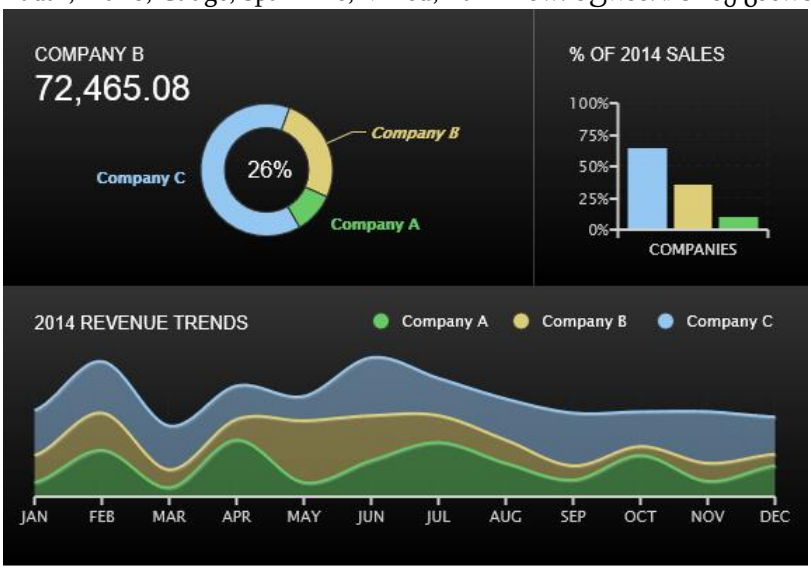
გამოსახატად. მუშაობს ყველა ბრაუზერზე. იხ. სურათი 4.



სურათი 4. Dygraph

5.2. ZingChart

ZingChart არის მძლავრი ბიბლიოთეკა მას აქვს შესაძლებლობა შექმნას გრაფიკები, დაშბორდები და ინფოგრაფიკები. გვთავაზობს გრაფიკების ასობით ვარიაციას და მეთოდს, როგორიცაა Bar, Scatter, Radar, Piano, Gauge, Sparkline, Mixed, Rank flow. სურათი 5-ზე გამოსახულია ZingChart.

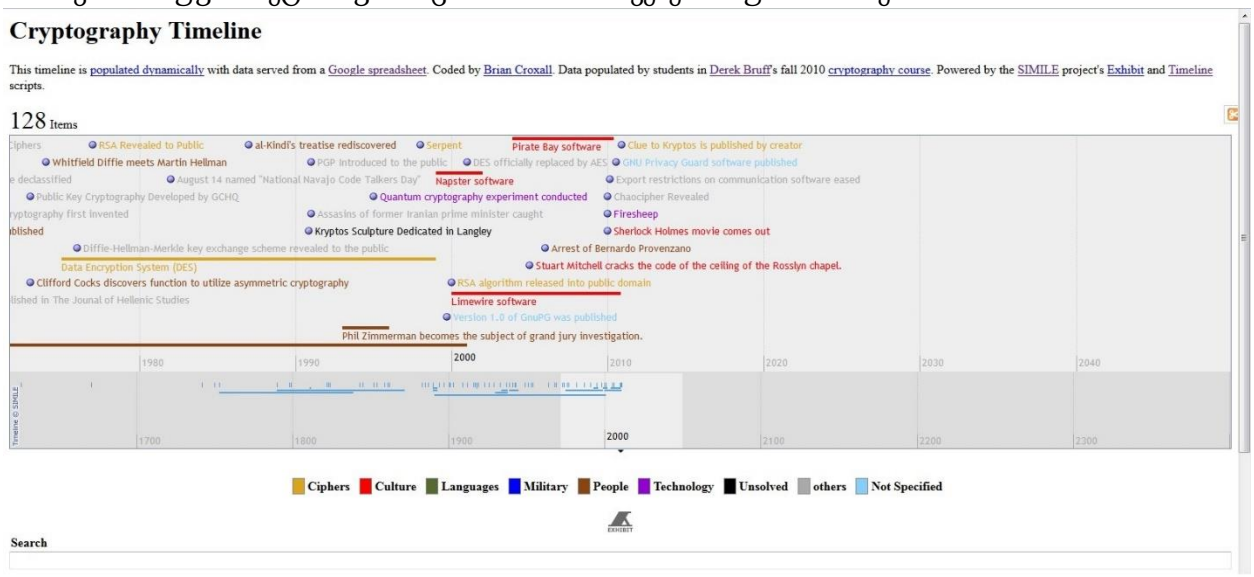


სურათი 5. ZingChart

5.3.Timeline

Timeline არის განსხვავებული ხელსაწყო, რომელიც არის ინტერაქტიული ტიპის, გვაწვდის დიდ ინფორმაციას კომპრესირებულ სივრცეში. შეგვიძლია დავაკლიკოთ თითოეულ ელემენტზე

რათა ვნახოთ უფრო მეტი ინფორმაცია. Timeline ნაჩვენებია სურათი 7-ზე.



სურათი 7. Timeline

6. შეჯამება და კვლევის ტენდენციები

ნაშრომში განვიხილეთ დიდ მონაცემებთან სამუშაოდ გამოყენებული ტექნიკები, მეთოდები, კონცეფტები და ხელსაწყოები. ასევე აღვწერეთ თუ როგორ ხდება დიდი რაოდენობის უწყვეტი მედია ნაკადების გადაცემა დამუშავება და შენახვა. დღეს მკვლევარები აქტიურად ეძებენ მეთოდებს რომ მეტად ოპტიმალურად მოხდეს დიდი მონაცემების ანალიზი, ძებნა, ვიზუალიზაცია. ეს მეთოდები უნდა სინქრონიზირდეს როგორც ინფორმაციულ ტექნოლოგიებთან, ასევე ბიზნესთან და რა თქმა უნდა იყოს უსაფრთხო და დაიცვას ინფორმაცია. დიდი მონაცემების ანალიზი კონცენტრირებულია ხელსაწყოებზე, ალგორითმებზე და არქიტექტურაზე.

ლიტერატურა

[1] Neelam Singh, Neha Garg, Varsha Mittal, Data – insights, motivation and challenges, Volume 4, Issue 12, December-2013, 2172, ISSN 2229-5518 2013.

[2] Karthik Kambatlaa, Giorgos Kollias b, Vipin Kumarc, Ananth Gramaa, Trends in big data Analytics, (2014) 74 2561–2573

[3] Francis X. “On the Origin(s) and Development of the Term \Big Data"_ Francis X., 2012

[4] Venkata narasimha inukollu1, sailaja arsi1 and srinivasa rao ravuri3 Security issues associated with big data in cloud computing Vol.6, No.3, May 2014

[5] Daniel Keim Big-Data Visualization.

[6] Hsinchun Chen Business Intelligence And Analytics: From Big Data To Big Impact AZ 85721, OH 45221-0211 U.S.A. Mack Robinson, GA 30302-4015.

[7] Edd Dumbill, Making Sense of Big Data

[8] Tackling the Challenges of Big Data 2014.

[9] http://fellinlovewithdata.com/research/the-role-of-algorithms-in-data-visualization