ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Լալայան Արթուր Գագիկի

ԵՐԿՐԻ ԴԻՏԱՐԿՄԱՆ ՏՎՅԱԼՆԵՐԻ ՀԱՄԱՐ ԱՄՊԱՅԻՆ ԵՎ ԲԱՐՁՐ ԱՐՏԱԴՐՈՂԱԿԱՆՈՒԹՅԱՄԲ ՀԱՐԹԱԿԻ ՄՇԱԿՈՒՄԸ

Ե.13.04 – «Հաշվողական մեքենաների, համալիրների, համակարգերի և ցանցերի մաթեմատիկական և ծրագրային ապահովում» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի համար

ՍԵՂՄԱԳԻՐ

Երևան 2023

INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS OF THE NAS RA

Lalayan Arthur

DEVELOPMENT OF A CLOUD AND HIGH-PERFORMANCE PLATFORM FOR EARTH OBSERVATION DATA

ABSTRACT

Of the dissertation for obtaining a Ph.D. degree in Technical Sciences on specialty 05.13.04 “Mathematical and Software Support of Computers, Complexes, Systems and Networks”

Yerevan 2023

Ատենախոսության թեման հաստատվել է Հայաստանի ազգային պոլիտեխնիկական համալսարանում

Գիտական ղեկավար՝	տեխ. գիտ. դոկտոր Հ. Վ. Ասցատրյան
Պաշտոնական ընդդիմախոսներ՝	ֆիզ. մաթ. գիտ. դոկտոր Ս. Կ. Շուքուրյան տեխ. գիտ. թեկնածու Ա. Մ. Բելոցերկովսկի
Առաջատար կազմակերպություն՝	«Հիդրոօդերևութաբանության և մոնիթորինգի կենտրոն» պետական ոչ առևտրային կազմակերպություն

Ատենախոսության պաշտպանությունը տեղի կունենա 2023թ. դեկտեմբերի 25-ին ժամը 15:00-ին ՀՀ ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող 037 «Ինֆորմատիկա» մասնագիտական խորհրդի նիստում հետևյալ հասցեով՝ Երևան, 0014, Պ. Սևակի 1:

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ ԻԱՊԻ գրադարանում:

Սեղմագիրն առաքված է 2023թ.-ի նոյեմբերի 24-ին:

Մասնագիտական խորհրդի գիտական քարտուղար ֆիզ. մաթ. գիտ. դոկտոր՝ Մ. Ե. Հարությունյան

The topic of the dissertation was approved at the National Polytechnic University of Armenia

Scientific supervisor:	H. V. Astsatryan, Ph.D., Doctor of Sciences
Official opponents:	S. K. Shoukourian, Ph.D., D.Ph.M.S. A. M. Belotserkovsky, Ph.D.
Leading organization:	“Hydrometeorology and Monitoring Center” State Non-Commercial Organization

The Defense will take place on December 25, 2023; at 15:00, at the Specialized Council 037 «Informatics» at the Institute of Informatics and Automation Problems of NAS RA. Address: Yerevan, 0014, P. Sevak 1.

The Dissertation is available in the library of IIAP NAS RA.

The abstract is delivered on November 24, 2023.

Scientific Secretary of the Specialized Council, D.Ph.M.S. M. E. Haroutunian

Աշխատանքի ընդհանուր նկարագիրը

Թեմայի արդիականությունը. Երկրի դիտարկման (ԵԴ) տվյալները ներկայացնում են արբանյակներից, ինքնաթիռներից, անօդաչու թռչող սարքերից և ցամաքային տվիչներից հավաքված տեղեկատվության հսկայական քանակությունը[1]: Այդպիսի տվյալները կարևոր են շրջակա միջավայրի մշտադիտարկման համար, քանի որ տեղեկատվություն են տրամադրում Երկրի աշխարհագրական թաղանթների մասին, ինչպիսիք են մթնոլորտը կամ ջրային ռեսուրսները[2]: ԵԴ տվյալները տրամադրում են շրջակա միջավայրի երկարաժամկետ փոփոխությունները դիտարկելու համար ժամանակային շարքի տեղեկատվություն, որոնց բարդությունն ու ծավալն աճում են՝ ստեղծելով դժվարություններ պահպանման, կառավարման և մշակման մակարդակներում:

ԵԴ տվյալների ծավալի աճը և բազմազանությունը, որը գնալով դառնում է ավելի բարդ, պահանջում է զգալի հաշվողական ռեսուրսներ և ճկուն ծրագրակազմ: Նման տվյալների մշակման աճող կարիքները բավարարելու համար սովորաբար օգտագործվում են բարձր արտադրողականությամբ հաշվողական (high-performance computing) ռեսուրսներ, որոնք տրամադրում են տվյալների կենտրոնները կամ ամպային մատակարարները:

Հետազոտական համայնքները ԵԴ տվյալների մշակման համար օգտագործում են ԵԴ մասնագիտացված հարթակներ կամ հաշվողական ռեսուրսների մատակարարների կողմից առաջարկվող ընդհանուր ծառայություններ և ենթակառուցվածքներ, ինչպիսիք են Amazon-ը, Google-ը կամ Microsoft-ը, որոնք տրամադրում են հաշվողական և պահեստավորման հսկայական հնարավորություններ, ինչպես նաև գլոբալ ԵԴ պահոցներ: Երկու մոտեցումներն էլ ունեն իրենց առավելություններն ու սահմանափակումները: ԵԴ տվյալների հասանելիության, մշակելու և վիզուալիզացիայի համար տարբեր մասնագիտացված հարթակներ[3] առաջարկում են համապարփակ լուծումներ, ինչպիսիք են Sentinel Hub (SH)[4], Google Earth Engine (GEE)[5], WEkEO[6], CREODIAS[7] և այլն: Նշված հարթակները փոխկապակցված են մատակարարների կողմից տրամադրվող հաշվողական ենթակառուցվածքների հետ: Մասնավորապես SH-ը տեղակայված է Amazon Web Services-ում, GEE-ն՝ Google Cloud Platform-ում, իսկ Copernicus-ի նախաձեռնությամբ WEkEO և CREODIAS ամպային հարթակները՝ CloudFerro-ում: Ուստի նման լուծումները ընդհանուր են և հասանելի են միայն այդ ենթակառուցվածքների օգտատերերին, որոնք վճարում են օգտագործված ռեսուրսների համար: Մատակարարի արգելափակման (vendor lock-in) խնդիրը ևս զգալի

[1] S.D. Jawak, V. Pohjola, et al. Status of Earth Observation and Remote Sensing Applications in Svalbard. Remote Sensing, 15(2):513, 2023.
[2] Q. Zhao, L. Yu, et al. An Overview of the Applications of Earth Observation Satellite Data: Impacts and Future Trends. Remote Sensing, 14(8):1863, 2022.
[3] V. C. F. Gomes, R. Q. Gilberto, R. F. Karine. An Overview of Platforms for Big Earth Observation Data Management and Analysis. Remote Sensing, 12(8):1253, 2020.
[4] Sentinel Hub-ի կայքէջն է. https://www.sentinel-hub.com/
[5] M. Amani, et al. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, pp. 5326-5350, 2020.
[6] WEkEO-ի կայքէջն է. https://www.wekeo.eu/
[7] CREODIAS-ի կայքէջն է. https://creodias.eu/

սահմանափակում է, քանի որ այլ հարթակի կամ ամպային մատակարարի անցնելը սեփական ձևաչափերի և գործիքների պատճառով դժվար է և թանկ: Հարկ է նշել, որ openEO-ն[8] տրամադրում է ԵԴ հարթակներից (ինչպիսիք են SH-ը, GEE-ը) օգտվելու ընդհանուր միջերեսներ, որը սակայն ինչպես վեբ ծառայություն չի կարող ապահովել արտադրողականության կամ ընդլայնելիության արդյունավետություն:

Մատակարարներից և ընդհանուր լուծումներից կախվածությունը հաղթահարելու համար սովորաբար կիրառվում է ԵԴ տվյալների խորանարդ (EODC)[9] հայեցակարգի բաց կոդով իրականացում Open Data Cube (ODC)[10] հարթակը, որը տվյալները պահպանում է խորանարդի ձևով՝ հեշտացնելով տարածական և ժամանակային վերլուծությունները: Որպես ինքնուրույն և հաշվողական ենթակառուցվածքներից անկախ հարթակ՝ ODC-ն ապահովում է ճկունություն, թույլ տալով օգտագործողներին ստեղծել սեփական օրինակները և հարմարեցնել համակարգը իրենց հատուկ պահանջներին: Սակայն ODC միջավայրում հորիզոնական ընդլայնելիություն, արտադրողականություն և այլ գործոններ հաշվի առնված չեն, որոնք կարևոր են լայնածավալ տվյալների արդյունավետ մշակման համար, իսկ երրորդ կողմի (third-party) հաշվողական ենթակառուցվածքների ավտոմատ տրամադրումը բարդ է:

Հաշվի առնելով վերոնշյալ սահմանափակումները՝ անհրաժեշտ է մշակել հաշվողական ենթակառուցվածքից անկախ ԵԴ տվյալների մշակման համալիր համակարգ, որը ապահովում է ճկուն և ընդարձակելի լուծումներ՝ հաշվի առնելով արտադրողականության կարևորագույն հիմնական ցուցանիշները:

Աշխատանքի նպատակը և դիտարկված խնդիրները: Աշխատանքի հիմնական նպատակն է մշակել հաշվողական ենթակառուցվածքից անկախ ԵԴ տվյալների մշակման ընդլայնվող համալիր համակարգ, որը համատեղում է տվյալների պահոցները ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների հետ: Այս նպատակին հասնելու համար դիտարկենք հետևյալ խնդիրները.

1. Մշակել ընդլայնվող, առանց սերվերի, տվյալների պահոցների ու ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների հետ փոխգործունակ, հաշվողական ենթակառուցվածքից անկախ ԵԴ տվյալների մշակման համալիր համակարգ, որն ապահովում է տվյալների արդյունավետ և ճկուն մշակում:
2. Մշակել ԵԴ տվյալների մշակման համար բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ օպտիմալացման մեթոդ, որը կապահովի ըստ պահանջի հաշվողական ռեսուրսների ընդլայնում՝ հաշվի առնելով տարբեր չափանիշներ, ինչպիսիք են արտադրողականությունը և ծախսերը:
3. Գնահատել մեծածավալ ԵԴ տվյալների բաշխված մշակման վրա տվյալների սեղմման մեթոդների ազդեցությունը՝ հավասարակշռություն հաստատելով պահեստավորման խնայողության և մշակման արագության բարելավման միջև:

[8] M. Schramm, E. Pebesma, et al. The openEO API–Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities. Remote Sensing, 13:1125, 2021.
[9] G. Giuliani, B. Chatenoux, et al. Data Cube on Demand (DCoD): Generating an earth observation Data Cube anywhere in the world. International Journal of Applied Earth Observation and Geoinformation, 87: 102035, 2020.
[10] Open Data Cube-ի դոկումենտացիան հասանելի է. https://www.opendatacube.org/.

Հետազոտության օբյեկտները: Այս աշխատությունում հետազոտության հիմնական օբյեկտներն են.

- ԵԴ տվյալների մշակման մեթոդներ, ալգորիթմներ և մոտեցումներ, որոնք կիրառվում են տվյալների վերլուծության համար:
- Բարձր արտադրողականությամբ հաշվարկներ, ամպային և մեծածավալ տվյալների մշակման հարթակներ, տեխնոլոգիական ենթակառուցվածքներ՝ մեծածավալ ԵԴ տվյալների արդյունավետ մշակման համար:
- ԵԴ տվյալների հետ կիրառվող տարբեր սեղմման մեթոդներ:

Հետազոտության մեթոդներ: Հետազոտությունում կիրառվել են բազմահոսք, բաշխված հաշվողական ծրագրավորում, բազմաֆունկցիոնալ օպտիմալացում, ռեգրեսիոն վերլուծություն, տվյալների միաձուլման մեթոդ, ամպային հաշվարկներ, առանց սերվերի ճարտարապետություններ, աշխարհատարածական վերլուծության համար Python և Java ծրագրավորման լեզուներ, ԵԴ տվյալների մշակման գրադարաններ, արտադրողականության ուսումնասիրություն ու գնահատում, և տվյալների բաշխված մշակման միջավայրեր (Hadoop, Spark և Dask):

Աշխատանքի գիտական նորույթը: Այս աշխատության համատեքստում ներկայացվում են հետևյալ գիտական արդյունքները.

1. Հաշվողական ենթակառուցվածքից անկախ ԵԴ տվյալների մշակման ընդլայնվող համալիր համակարգ, որը հաշվի է առնում միջազգային հիմնօրինակները և բավարարում է մեծածավալ տվյալների արդյունավետ մշակման և պահպանման համար դիտարկված հիմնական կատարողական ցուցանիշներին:
2. ԵԴ տվյալների արդյունավետ մշակման համար բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ մեթոդ, որը հաշվի է առնում հաշվողական ենթակառուցվածքների առանձնահատկությունները և աշխատանքային հոսքերի բարդությունը:
3. ԵԴ տվյալների պահպանման համար նախատեսված արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն, որը տվյալների մշակման արտադրողականության բարձրացման համար առաջարկում է տվյալների սեղմման արդյունավետ մեթոդներ:

Ստացված արդյունքների կիրառական նշանակությունը: Մշակված համալիր համակարգը կարող է օգտագործվել լայնածավալ ԵԴ տվյալների արդյունավետ մշակման համար՝ հաշվի առնելով տվյալների մշակման արտադրողականության և ծախսերի գործոնները, օգտագործելով ամպային կամ բարձր արտադրողականությամբ ենթակառուցվածքներ:

Ներդրումներ: Մշակված համակարգը ներդրվել է «ՖՈՐԵՍԹԲԵՌԳ» ՍՊԸ-ում և օգտագործվում է անտառային միջավայրերի մշտադիտարկման համար՝ ապահովելով ԵԴ տվյալների արդյունավետ և արագ մշակում:

Պաշտպանության ներկայացվող հիմնական դրույթները:

1. ԵԴ տվյալների մշակման համալիր համակարգ, որը հաշվողական ենթակառուցվածքից անկախ է և ճկուն կերպով միավորում է ԵԴ տվյալների պահոցները տեղական կամ գլոբալ ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների հետ:
2. Բաշխված հաշվողական կլաստերի ընտրության համար բազմաֆունկցիոնալ օպտիմալացման մեթոդ, որն ապահովում է ԵԴ տվյալների արդյունավետ մշակումը՝ միաժամանակ հաշվի առնելով արտադրողականությունը և ծախսարդյունավետությունը:
3. Արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն, որն առաջարկում է տվյալների սեղմման արդյունավետ մեթոդներ՝ բարձրացնելով ԵԴ տվյալների բաշխված մշակման արտադրողականությունը:

Ստացված արդյունքների գրաքննությունը և փորձարկումը: Ստացված արդյունքները զեկուցվել են միջազգային մի շարք գիտաժողովներում.

1. 13th Conference on Data Analysis Methods for Software Systems (DAMSS), Druskininkai, Lithuania, December 1-3, 2022,
2. 14th International Conference on Large-Scale Scientific Computations (LSSC), Sozopol, Bulgaria, June 5-9, 2023,
3. 14th International Conference on Computer Science and Information Technologies (CSIT), Yerevan, Armenia, September 25-30, 2023.

Աշխատանքի արդյունքները քննարկվել են Հայաստանի ազգային պոլիտեխնիկական համալսարանում, ՀՀ ԳԱԱ ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում անցկացված սեմինարների ընթացքում:

Հրապարակումներ: Ատենախոսության հիմնական արդյունքները հրապարակվել են յոթ (7) գիտական աշխատություններում (4-ը WoS/Scopus-ում), որոնց ցանկը բերված է սեղմագրի վերջում:

Աշխատանքի ծավալը և կառուցվածքը: Ատենախոսության ծավալը կազմում է 109 էջ, ներառում է 124 գրականության հղում և բաղկացած է ներածությունից, 4 գլուխներից և օգտագործված գրականության ցանկից:

Աշխատանքի բովանդակությունը

Ներածություն բաժնում հիմնավորվում է ատենախոսության արդիականությունը, ձևակերպված է աշխատանքի նպատակը, դիտարկված խնդիրները, գիտական նորույթը, կիրառական նշանակությունը և պաշտպանության ներկայացված հիմնական դրույթները:

Առաջին գլխում ներկայացվում է աշխատանքի ներածությունը և նշանակությունը: Գլուխը ներառում է նաև հաշվողական ենթակառուցվածքները, հարթակները և ծառայությունները, հիմնօրինակները, գործիքները և տվյալների

ձևաչափերը, որոնք անհրաժեշտ են լայնածավալ ԵԴ տվյալների արդյունավետ կառավարման և մշակման համար:

1.1 ենթագլխում ներկայացված են ԵԴ տվյալների ներածությունը և արդյունավետ մշակման կարևորությունը:

1.2 ենթագլխում նկարագրված է ԵԴ տվյալների նշանակությունը շրջակա միջավայրի մշտադիտարկման համար: Այս ենթագլուխն ուսումնասիրում է ԵԴ տվյալների հետ աշխատանքին բնորոշ առավելություններն ու մարտահրավերները: Ենթագլխում բերված է համապարփակ վերլուծություն ԵԴ բաց տվյալներ տրամադրող արբանյակների մասին՝ պարզաբանելով դրանց բնութագրերը, տվիչային գոտիները և ինդեքսները, որոնք հեռահար զոնդավորման տեղեկատվությունից ստացված հաշվարկներ կամ ալգորիթմներ են նախատեսված բնապահպանական պայմանների մասին տեղեկատվություն ստանալու համար:

1.3 ենթագլխում ներկայացված են առանց սերվերի, ամպային և բարձր արտադրողականությամբ հաշվողական ենթակառուցվածքները, որոնք արդյունավետորեն կիրառվում են ԵԴ տվյալները կառավարելու համար: Նաև ներկայացված է բաշխված մշակման հայեցակարգը՝ վերլուծելով մեծածավալ տվյալների հավաքածուները բաշխված և զուգահեռ մշակելու համար լայնորեն կիրառվող բաց կոդով Apache Hadoop[11], Spark[12] և Dask[13] միջավայրերը:

Այս ենթագլուխը նկարագրում է նաև ԵԴ գլոբալ պահոցները, ինչպես նաև լայնորեն կիրառվող գլոբալ ամպային մատակարարների կողմից տրամադրվող հայտնի լուծումները, ԵԴ տվյալների մասնագիտացված հարթակներն ու ծառայությունները: Ենթագլուխը գնահատում է լուծումների առավելությունները և սահմանափակումները: Մասնավորապես ներկայացվում է հաշվողական ենթակառուցվածքից անկախ և բաց կոդով տվյալների խորանարդ հարթակի հնարավորությունները, որոնք պարզեցնում են մեծածավալ ԵԴ տվյալների մշակումն ու վերլուծությունը: Հատկանշական է, որ ԵԴ տվյալների կառավարման և վերլուծության համար վերոնշյալ հարթակը ներդրվել և օգտագործվել է տարբեր երկրներում, այդ թվում՝ Ավստրալիայում, Շվեյցարիայում, Բրազիլիայում և Հայաստանում: Բնապահպանական տարբեր մարտահրավերներին դիմակայելու և գլոբալ մշտադիտարկման ջանքերը հեշտացնելու համար ներկայացված են հարթակի կիրառման արդյունավետությունը:

1.4 ենթագլուխը նվիրված է աշխարհատարածական տվյալների համար հատուկ մշակված միջազգային հինօրինակներին, ինչպես նաև ԵԴ տվյալների համար հասանելի ձևաչափերին: Այս ենթագլուխը ներկայացնում է նոր տվյալների ձևաչափերի և ԵԴ տվյալների կառավարման համար ստեղծված գործիքների

[11] T. Hussain, A. Sanga, et al. Big data hadoop tools and technologies: A review. Proceedings of International Conference on Advancements in Computing & Management (ICACM). 2019.
[12] E. Shaikh, et al. Apache spark: A big data processing engine. 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM). IEEE, 2019.
[13] M. Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. Proceedings of the 14th python in science conference, 130, 2015.

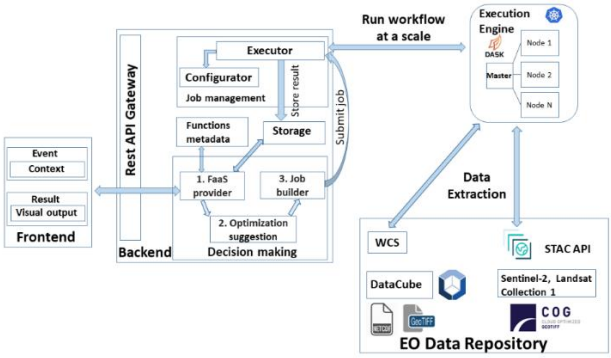
օգտագործման նպատակահարմարությունը, որոնք առանցքային դեր են խաղում՝ ապահովելով արդյունավետ տվյալների մշակում, պահպանում և անխափան հասանելիության տրամադրում:

1.5 ենթագլուխը եզրափակում է սույն գլուխը՝ տրամադրելով ԵԴ տվյալների կառավարման և մշակման համար օգտագործվող մեթոդների համապարփակ ակնարկ, դրանց արդյունավետությունը և լուծումների սահմանափակումները: Ներկայացված են հետազոտության մարտահրավերները, նպատակները և դիտարկված խնդիրները, որոնք հնարավորություն կտան հաղթահարել այդ սահմանափակումները:

Երկրորդ գլխում ներկայացված է առաջարկվող ընդլայնվող ԵԴ տվյալների մշակման համալիր համակարգը[14]:

2.1 ենթագլուխը ներկայացնում է ԵԴ տվյալների արդյունավետ մշակման առկա լուծումների մանրակրկիտ հետազոտություն: Առաջարկվող լուծումների արդյունավետությունը գնահատելու համար օգտագործվում են հիմնական կատարողական ցուցանիշներ, ինչպիսիք են արտադրողականությունը և ընդլայնելիությունը, ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների և ԵԴ տվյալների պահոցների հետ փոխգործունակությունը, ավտոմատ և արագ ամպային և բարձր արտադրողականությամբ ռեսուրսների տրամադրումը և ընդլայնումը, նոր ձևաչափերի և լուծումների աջակցումը, կապը բաց կոդով ծրագրային միջավայրերի հետ, ինչպես նաև ODC-ի հետ: Ներկայացված է այլ հեղինակների կողմից առաջարկվող լուծումների սահմանափակումների գնահատումը ըստ կատարողական ցուցանիշների:

2.2 ենթագլխում ներկայացված է առաջարկվող համակարգը, որի ճարտարապետությունը բերված է Նկար 1-ում:



Նկար 1: ԵԴ տվյալների մշակման ընդլայնվող համալիր համակարգի կառուցվածքը

[14] Կոդը հասանելի է https://github.com/ArmHPC/Scalable-EO-system

Համակարգը բաղկացած է 4 հիմնական մոդուլներից՝ Frontend, Backend, Execution engine և EO data repositories: Առանձնացնելով տվյալների պահոցները հաշվողական ենթակառուցվածքներից՝ համակարգը ապահովում է ճկուն փոխգործունակություն, աջակցելով տվյալների պահպանման տարբեր ձևաչափերին: Frontend մոդուլը ապահովում է օգտագործողին հարմար միջերես՝ թաքցնելով ԵԴ տվյալների մշակման բարդությունները, մինչդեռ Backend-ը, որը ներառում է որոշումների կայացման և աշխատանքի կառավարման բաղադրիչներ, մշակում է RESTful API-ի միջոցով ստացված հարցումները, հաշվի առնելով տարածքը, ժամանակահատվածը և մշակման ֆունկցիայի պարամետրերը: Համակարգը նախագծված է ճկուն կերպով, որը թույլ է տալիս հեշտությամբ ընդլայնել՝ ներդնելով ԵԴ տվյալների մշակման նոր գործառույթներ: Dask-ի վրա հիմնված Execution engine մոդուլը օգտագործվում է տվյալների բաշխված մշակման համար, որը հաշվողական ենթակառուցվածքից անկախ է, ապահովում է համակարգի ընդլայնելիությունը և արտադրողականությունը՝ օգտագործելով տեղական կամ գլոբալ ամպային կամ բարձր արտադրողականությամբ հաշվողական ենթակառուցվածքներ: EO data repositories մոդուլը տրամադրում է ԵԴ տվյալների հասանելիություն՝ ապահովելով պահոցների հետ փոխգործունակություն, համակարգին դարձնելով ավելի ճկուն և ընձեռնելով հնարավորություն աշխատել տարբեր պահոցների հետ: Այս ենթագլխում մանրամասն նկարագրվում է աշխատանքի ընթացքը, կազմաձևման պարամետրերը և գնահատումը վեբ հավելվածի միջոցով՝ ցուցադրելով արդյունավետությունը նորմալացված տարբերության բուսականության ինդեքսի[15] (Normalized Difference Vegetation Index - NDVI) կիրառմամբ, որը բուսականության խտությունը որոշելու գրաֆիկական ցուցիչ է: Փորձերը ցույց են տալիս հաշվողական ռեսուրսների չափման և մշակման արագության օպտիմալացման հարցում համակարգի արդյունավետությունը՝ հաշվի առնելով մուտքային տվյալների ծավալը:

2.3 ենթագլխում ներկայացվում է համակարգի հնարավորությունները, որպես ծառայություն օգտագործելով օդի աղտոտվածության մշտադիտարկումը՝ հատկապես թիրախավորելով ազոտի երկօքսիդի (NO_2) մակարդակները ընտրված տարածքների համար:

2.4 ենթագլխում ամփոփվում են 2-րդ գլխում ստացված արդյունքները:

Երրորդ գլուխը ներկայացնում է ԵԴ տվյալների մշակման աշխատանքային հոսքերի համար բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ մեթոդ, որը հնարավորություն է տալիս ընտրել տվյալների բաշխված մշակման արդյունավետ հաշվողական ռեսուրսներ: Մեթոդը կենտրոնացած է տվյալների մշակման արտադրողականության և ծախսարդյունավետության միջև հավասարակշռություն հաստատելու վրա՝ հաշվի

[15] D. Montero, C. Aybar, M. D. Mahecha et al, A standardized catalogue of spectral indices to advance the use of remote sensing in Earth system research. Scientific Data, 10: 197, 2023.

առնելով ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների առանձնահատկությունները և աշխատանքային հոսքերի բարդությունները:

3.1 ենթագլխում ներկայացվում է խնդրի ձևակերպումը՝ ընդգծելով ԵԴ տվյալների մշակման աշխատանքային հոսքերում բազմաֆունկցիոնալ օպտիմալացման նշանակությունը: ԵԴ տվյալների արդյունավետ մշակման առկա աշխատանքները վերաբերվում են տվյալների մշակման որակի բարելավմանը, ինչպես օրինակ դիտարկումների ընտրությանը և պլանավորմանը՝ օգտագործելով տվյալների բաշխված մշակման միջավայրեր, ռեգրեսիոն մոդելներ կամ գենետիկ ալգորիթմներ:

Թեև ԵԴ տվյալների մշակման արդյունավետությունը շատ կարևոր է, այն պետք է հավասարակշռված լինի հաշվողական ռեսուրսների օգտագործման տնտեսական արդյունավետության հետ, հատկապես ամպի վրա հիմնված միջավայրերում, որտեղ ծախսերը կապված են ռեսուրսների սպառման հետ: Տվյալների բաշխված մշակման համար արդյունավետ կլաստերի կազմաձև ընտրելիս արտադրողականության և արժեքի միջև արդյունավետ փոխզիջում գտնելը դժվար է, քանի որ մի կողմից ավելի շատ հաշվողական ռեսուրսների օգտագործումը կարող է հանգեցնել տվյալների մշակման ավելի մեծ զուգահեռացմանը և ժամանակի կրճատմանը, իսկ մյուս կողմից այն ավելի մեծ ծախսեր է առաջացնում ամպային մատակարարների կողմից: Բացի այդ, արդյունավետ կլաստերի կազմաձևի ընտրությունը NP դասի խնդիր է:

ԵԴ տվյալների մշակման առաջադրանքի արտադրողականության և ծախսերի նպատակները կարելի է ներկայացվել հետևյալ բանաձևերով.

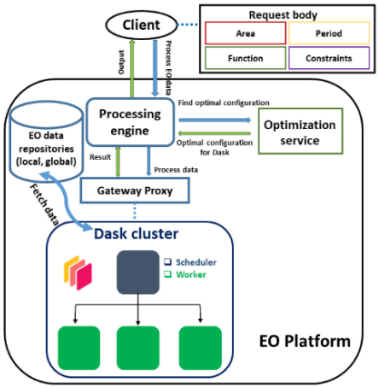
$$\mathrm{t} = \tau(\mathrm{s}, \mathrm{n}, \mathrm{r})$$
$$\mathrm{p} = \upsilon(\mathrm{t}, \mathrm{n}, \mathrm{r})$$
$$\mathrm{r} \in \mathrm{R};\ \mathrm{n}, \mathrm{s} \in \mathrm{N}$$

որտեղ τ-ն և υ-ն համապատասխանաբար արտադրողականության և ծախսերի նպատակային գործառույթներն են: t-ն ԵԴ առաջադրանքի կատարման ժամանակն է, հաշվի առնելով s բարդությունը կախված մուտքային տվյալների ծավալից և առաջադրանքի բարդությունից: Հաշվողական կլաստերը բաղկացած է n թվով հանգույցներից, որոնցից յուրաքանչյուրն ունի r տեսակի հաշվողական հանգույցներ դիտարկված R վերջավոր շարքից: ԵԴ տվյալների մշակման առաջադրանքի p արժեքը կախված է առաջադրանքի կատարման ժամանակից և հաշվողական կլաստերի բնութագրերից, մասնավորապես հանգույցների քանակից և հանգույցների տեսակից: Ստորև բերված բանաձևն օգտագործվում է հաշվողական ռեսուրսների արդյունավետ համակցությունը գտնելու համար՝ հաշվի առնելով արտադրողականության և ծախսերի նպատակները.

$$\min_{r \in R} [\mathrm{t} = \tau(\mathrm{s}, \mathrm{n}, \mathrm{r}), \mathrm{p} = \upsilon(\mathrm{t}, \mathrm{n}, \mathrm{r})]$$
$$0 < t \leq t', 0 < p \leq p'$$

որտեղ t'-ն ու p'-ն առաջադրանքի կատարման ժամանակի և ծախսերի բյուջեի սահմանափակումներն են:

3.2 Ենթագլուխը ներկայացնում է ԵԴ տվյալների մշակման համալիր համակարգում բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ մեթոդի իրականացումը հաշվողական ռեսուրսների արդյունավետ և ըստ պահանջի ընդլայման համար (տես Նկար 2):



Նկար 2: Արդյունավետ հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ մեթոդի կիրառման աշխատանքային հոսքի գծապատկերը

Համակարգում մեթոդը հանդիսանում է որպես ծառայություն՝ հնարավորություն տալով ստեղծել արդյունավետ կլաստերային կազմաձևեր հաշվի առնելով օգտագործողների կարիքները, մասնավորապես ԵԴ տվյալների մշակման արտադրողականությունը և ծախսարդյունավետությունը: Մեթոդի կիրառումը նկարագրվում է հետևյալ աշխատանքային հոսքով.

1. Օգտագործողը հարցում է ներկայացնում Processing engine-ին տրամադրելով անհրաժեշտ պարամետրերը, ինչպիսիք են հետաքրքրության տարածքը, ժամանակահատվածը, մշակման առաջադրանքը, կատարման ժամանակի (t') և ծախսերի (p') սահմանափակումները:
2. Այնուհետև հաշվողական կլաստերի արդյունավետ կազմաձևերը որոշելու համար Processing engine-ը հարցումն ուղարկում է օպտիմալացման ծառայություն:
3. Processing engine-ն օգտագործում է օպտիմալացման ծառայության առաջարկած արդյունավետ կազմաձևերը Dask gateway-ի միջոցով կլաստեր ստեղծելու համար: Լավագույն կատարումն ապահովող կազմաձևն ընտրվում է լռելյայն, եթե կան մի քանի արդյունավետ կազմաձևեր՝ հաշվի առնելով արտադրողականության և ծախսերի նպատակները միաժամանակ:
4. Processing engine-ը ստեղծում է հաշվողական գրաֆ՝ հաշվի առնելով հաճախորդի հարցումը, և իրականացնում այն Dask կլաստերում, որը պահանջվող տվյալները ներբեռնում է ԵԴ պահոցներից և սկսում տվյալների բաշխված մշակումը:
5. Վերջապես, մշակված արդյունքը առաքվում է օգտագործողին:

Մեթոդում Հայկական տվյալների խորանարդը[16] ծառայում է որպես տեղական, իսկ Sentinel-2 Cloud-Optimized GeoTIFF պահոցը[17] որպես գլոբալ ԵԴ տվյալների շտեմարան: Արդյունավետ կլաստերի ընտրության մեթոդը հիմնված է բազմաֆունկցիոնալ Պարետո օպտիմալացման մեթոդի[18] վրա, որն ընտրվել է տարբեր օպտիմալացման ալգորիթմների և մեթոդների, ներառյալ գենետիկական և էվոլյուցիոն ալգորիթմների համապարփակ վերլուծությունից հետո: Այն տրամադրում է համեմատաբար ավելի ճշգրիտ արդյունքներ և հատկապես նպատակահարմար է իր ճկունության և ընդլայնման հեշտության շնորհիվ, որը թույլ է տալիս պարզորեն ներառել լրացուցիչ նպատակներ, ինչպիսին է էներգիայի սպառումը: Մեթոդի ալգորիթմական տեսքը ներկայացված է ստորև.

Algorithm Optimization algorithm

Require: s, t', p' ▷ Task complexity, execution time and cost constraints
Ensure: $\min_{r \in R} \left[t = \tau(s,n,r), p = \upsilon(t,n,r) \right]$ subject to: $t \leq t', p \leq p'$
  $configs \leftarrow$ **finite set of Dask cluster configurations**
  $results \leftarrow \{\}$
  $optimalPoints \leftarrow \{\}$
  for $config$ **in** $configs$ **do**
    $time \leftarrow \tau(s,n,r)$ ▷ find or predict execution time for the given $config$ and complexity s
    $cost \leftarrow time \times config.instanceRate \times config.nodes$
    if $cost \leq p'$ **AND** $time \leq t'$ **then**
      $results$.append($(config, cost, time)$)
    end if
  end for
  for r **in** $results$ **do**
    $nonDominatedPoints \leftarrow \{r.cost > it.cost$ **AND** $r.time > it.time$ **for it in** $results\}$
    if $nonDominatedPoints$ **is empty then**
      $optimalPoints$.append(r)
    end if
  end for
  return $optimalPoints$

Առաջարկվող մեթոդը սկսվում է առաջադրանքի կատարման ժամանակի և պահանջվող հաշվողական ռեսուրսների արժեքի գնահատմամբ՝ հաշվի առնելով դիտարկվող վերջավոր հավաքածուից հաշվողական կլաստերի տարբեր կազմաձևեր: Տվյալ առաջադրանքի կատարման ժամանակի գնահատման գործընթացը՝ հաշվի առնելով τ կատարման նպատակային ֆունկցիան, ներառում է պատմական մոդելավորման տվյալների բազայի ուսումնասիրություն՝ ստուգելու համար, թե արդյոք նմանատիպ համեմատելի բարդությամբ առաջադրանք իրականացվել է, հաշվի առնելով մուտքային տվյալների ծավալը և դիտարկվող

[16] A. Asmaryan, V. Muradyan, et al. Paving the Way towards an Armenian Data Cube. Data 2019, 4, 117.
[17] Sentinel-2 Cloud-Optimized GeoTIFFs պահոցի կայքէջն է. https://registry.opendata.aws/sentinel-2-l2a-cogs/
[18] S. Petchrompo, D. W. Coit, et al. A review of Pareto pruning methods for multi-objective optimization, Computers & Industrial Engineering, 167:108022, 2022.

առաջադրանքի բարդությունը: Հակառակ դեպքում, նախապես ուսուցանված ռեգրեսիոն մոդելը կանխատեսում է տվյալների մշակման առաջադրանքի կատարման ժամանակը: Այնուհետև օպտիմալացման ծառայությունը հաշվարկում է յուրաքանչյուր կազմաձևի արժեքը՝ օգտագործելով υ նպատակային ֆունկցիան, բազմապատկելով կատարման գնահատված ժամանակը աշխատող հանգույցների քանակով և օգտագործվող հաշվողական հանգույցի ժամային դրույքաչափով: Եթե ստացված արժեքը և կատարման ժամանակը համընկնում են օգտագործողի տրամադրած սահմանափակումների հետ, ապա այն ավելացվում է հարցմանը բավարարող լուծումների ցանկում: Կազմաձևերի ստացված ցանկը զտվում է՝ պահպանելով այն տարրերը, որոնք գերազանցում են մյուսներին, միաժամանակ հաշվի առնելով ծախսերի և արտադրողականության նպատակները: Մասնավորապես, ավելի բարձր արժեքով և կատարման ժամանակով կազմաձևերը հանվում են ցանկից, որի արդյունքում ստացվում է այնպիսի կազմաձևերի ցանկը, որոնք առաջարկում են համեմատելի արժեք և արդյունավետություն, ինչպես նաև չունեն էական առավելություններ միմյանց նկատմամբ:

Մեթոդի արդյունավետությունը բարձրացնելու և ծախսերը կրճատելու համար անհրաժեշտ է կատարել զգալի թվով հաշվարկներ և փորձեր, որոնք ծախսատար են: Ենթագլուխը ներկայացնում է ԵԴ տվյալների մշակման աշխատանքային հոսքերի համար մշակված EO Cloud Simulator (EOCSim) մոդելը, որը հանդիսանում է առաջարկվող մեթոդի կարևոր մաս և հիմնված CloudSim[19] գործիքի վրա: Մոդելը հնարավորություն է տալիս վիրտուալ միջավայրում կատարել սիմուլյացիաներ, որոնց արդյունքները կիրառվել են առաջարկվող մեթոդի արդյունավետությունը բարձրացնելու նպատակով: Մոդելը տրամադրում է ամպային բարդ համակարգերի մոդելավորման և վերլուծության համար նախատեսված միջավայր և հիմնված է միլիոն հրահանգներ վայրկյանում (Million instructions per second-MIPS) հատկության վրա, որը թույլ է տալիս հեշտությամբ համեմատել տարբեր հաշվողական սարքերի մշակման հզորությունները: Մոդելի կիրառումը օգնում է գնահատել պրոցեսորների և համակարգերի հարաբերական աշխատանքը՝ հաշվի առնելով ոչ միայն պրոցեսորի արագագործությունը, այլ նաև ճարտարապետության առանձնահատկությունները՝ հնարավորություն տալով գնահատել և բարելավել ամպային ենթակառուցվածքների վրա հիմնված լուծումների կատարման արդյունավետությունը: Առաջարկված մոդելը ցուցադրում է իր արդյունավետությունը՝ տրամադրելով ԵԴ տվյալների բաշխված մշակման ժամանակի և ծախսերի գնահատում նկարագրված կլաստերում:

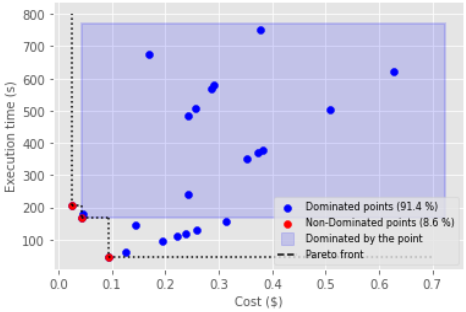
3.3 Ենթագլուխը գնահատում է առաջարկվող մեթոդը և քննարկում գնահատման արդյունքները: Գնահատումները կատարվել են օգտագործելով Ամերիկյան CloudLab[20] փորձարարական հաշվողական ենթակառուցվածքի և հայկական

[19] A. Sundas, S. N. Panda, et al. An Introduction of CloudSim Simulation tool for Modelling and Scheduling. 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 263-268, 2020.

[20] D. Duplyakin, R. Ricci, A. Maricq, et al. The Design and Operation of CloudLab. In Proceedings Of The USENIX Annual Technical Conference (ATC), 2019.

ամպային ենթակառուցվածքի[21] հաշվողական ռեսուրսները, որոնք առաջարկում են համայնքներին տարատեսակ ծառայություններ: Առաջարկվող արդյունավետ հաշվողական կլաստերի ընտրության մեթոդը ենթարկվել է գնահատման բազմաթիվ Dask կլաստերի կազմաձևերի վրա՝ ընդգրկելով աշխատող հանգույցների բազմազան տեսակները: Օգտագործվել են տարբեր քանակի հանգույցներ՝ մեկ միջուկից և 2 ԳԲ օպերատիվ հիշողություն ունեցող հաշվողական հանգույցից մինչև 64 միջուկ և 128 ԳԲ օպերատիվ հիշողությամբ հանգույց: Մեթոդը գնահատելու համար հետազոտությունը կենտրոնացած էր Հայաստանի տարածքի վերլուծության վրա՝ օգտագործելով Sentinel-2 արբանյակային տվյալների NIR, RED, BLUE և GREEN գոտիները: Դիտարկվել են երեք տարբեր աշխատանքային ծանրաբեռնվածություններ, որոնցից յուրաքանչյուրը տարբեր մուտքային տվյալների չափերով համապատասխանում է շաբաթական (թեթև - 0,08 ՏԲ), ամսական (միջին - 0,32 ՏԲ) և սեզոնային (ծանր 1,2 ՏԲ) ժամանակաշրջաններին:

Նկար 3-ը ցույց է տալիս Պարետո ճակատը շաբաթական ծանրաբեռնվածության համար՝ նշելով փոխզիջում մրցակցային արտադրողականության և ծախսերի նպատակների միջև:



Նկար 3: Պարետո ճակատը շաբաթական ծանրաբեռնվածության համար:

Առաջարկվող մեթոդը ցույց է տալիս, որ յուրաքանչյուր ծանրաբեռնվածության համար կազմաձևերի միայն չնչին տոկոսն է համարվում Պարետո-արդյունավետ: Մասնավորապես, սեզոնային ծանրաբեռնվածության համար կազմաձևերի միայն 5,7%-ն է Պարետո-արդյունավետ (երկու արդյունավետ կետեր՝ մեկը ապահովում է լավագույն կատարման ժամանակ, իսկ մյուսը՝ նվազագույն արժեք): Պարետո-արդյունավետ լուծումների քանակը կախված է ուսումնասիրված հաշվողական կլաստերի կազմաձևերի բարդությունից: Ի հակադրություն, շաբաթական և ամսական ծանրաբեռնվածության առկա տարբերակներից հայտնաբերվեցին միայն երեք Պարետո-արդյունավետ կազմաձևեր: Արդյունավետ կազմաձևերից որևէ մեկի ընտրությունը կարող է բարելավել արտադրողականությունը՝ միաժամանակ

[21] H. Astsatryan, V. Sahakyan, et al. Strengthening compute and data intensive capacities of Armenia. In 2015 14th RoEduNet International Conference - Networking in Education and Research (RoEduNet NER), 2015.

նվազեցնելով հաշվողական ռեսուրսների կիրառման համար պահանջվող արժեքը: Շաբաթական ծանրաբեռնվածությունը ուսումնասիրելիս, բոլոր դիտարկված կլաստերների կազմաձևերի կիրառման դեպքում կատարման միջին ժամանակի և արժեքի համեմատությունը արդյունավետ կազմաձևերի դեպքում ստացված միջին արժեքների հետ ցույց է տալիս, որ հնարավոր է հասնել միջինը 2,3 անգամ ավելի արագ կատարման ժամանակի և 4,7 անգամ կրճատված ծախսերի:

EOCSim մոդելի գնահատման արդյունքները ցուցադրում են, որ մոդելը ունի բարձր ճշգրտություն, երբ համեմատվում է իրական արդյունքների հետ, օրինակ՝ Հայաստանի տարածքի համար շաբաթական NDVI-ի աշխատաժամանակի կանխատեսման դեպքում R^2=0.88, իսկ միջին քառակուսային սխալանքը RMSE=78՝ հաշվի առնելով մի շարք կլաստերային կազմաձևեր, որոնցից յուրաքանչյուրն ունի տարբեր քանակի և տեսակի հանգույցներ:

3.4 ենթագլուխը հակիրճ ամփոփում է 3-րդ գլխում ստացված արդյունքները:

Գլուխ 4-ը հետազոտում է սեղմման տեխնիկայի ազդեցությունը լայնածավալ ԵԴ տվյալների հավաքածուների մշակման վրա՝ նպատակ ունենալով գտնել հավասարակշռություն պահեստային տարածքի խնայման և տվյալների մշակման արագության բարձրացման միջև: Փոխզիջման հայտնաբերումը հնարավորություն կտա խնայել պահոցի հիշողությունը՝ միաժամանակ պահպանելով կամ բարելավելով մշակման արդյունավետությունը:

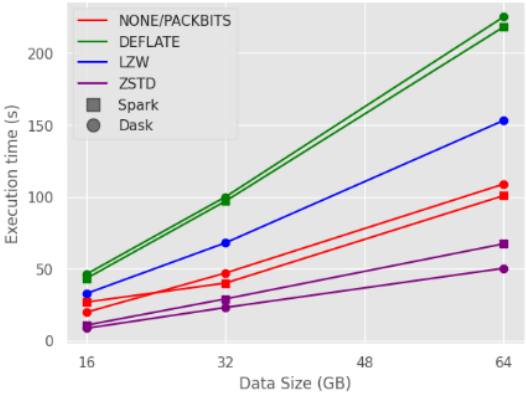
4.1 ենթագլխում ներկայացված է աշխատանքի նախապատմությունը, ներառյալ լայնածավալ տվյալների մշակման օպտիմալացման գոյություն ունեցող մոտեցումները: Պահեստային տարածքի պահպանման և տվյալների մշակման արագության փոխզիջմանը հասնելու համար ներկայացված են տվյալների սեղմման ազդեցությունը տվյալների բաշխված մշակման հայտնի միջավայրերի համար: Այլ հեղինակների աշխատանքներում հետազոտված են մեծածավալ տվյալների միջավայրերի արտադրողականության բարձրացումը որոշ սահմանափակումներով, որոնք ազդում են որոշումների կայացման վրա: Հեղինակների կողմից ցույց է տրված, որ միջավայրերի արտադրողականությունը բարձրանում է, երբ մուտքային տվյալները սեղմված են, կամ երբ փոփոխվել են որոշ կազմաձևման պարամետրեր, օրինակ կրկնօրինակների քանակը:

Մեր կողմից առաջարկվող լուծումը ներառում է տարբեր աշխատանքային հոսքերի համար սեղմման արդյունավետ մեթոդների հայտնաբերումը՝ հաշվի առնելով տվյալների մշակման ժամանակը: Դիտարկվել են տեքստային, թվային, ինչպես նաև արբանյակային պատկերների համար հասանելի առանց կորստի տվյալների սեղմման բոլոր մեթոդները, որոնք աջակցվում են մեծածավալ տվյալների մշակման միջավայրերի կողմից:

4.2 ենթագլխում ներկայացված են գնահատումներ, որոնք ընդգծում են տվյալների սեղմման արդյունավետությունը մեծածավալ ԵԴ տվյալների մշակման ժամանակ: Գնահատումները ցույց են տալիս, որ սեղմման մեթոդները ունեն սեղմման տարբեր գործակիցներ և ցուցադրում են յուրահատուկ վարքագիծ տվյալների բաշխված մշակման ընթացքում: Մեծածավալ տվյալների միջավայրերում

տվյալների սեղմման մեթոդների արդյունավետությունը գնահատելու համար ընտրվել են տեքստային/թվային ԵԴ տվյալների համար WordCount և LogAnalyzer հավելվածները, որոնք օգտագործվում են տվյալների վերլուծության և մշակման մեջ, ներառյալ տեքստային և թվային տվյալների զտումը, ինչպիսիք են հաշվետվությունները, մետատվյալները, և տվիչային տվյալները, մեքենայական ուսուցման K-Means ալգորիթմը, որը լայնորեն օգտագործվում է առանձնահատկությունների արդյունահանման, դասակարգման կամ անոմալիաների հայտնաբերման նպատակով և ԵԴ տվյալների մշակման NDVI ինդեքսը՝ արբանյակային նկարների մշակման համար:

Dask և Spark կլաստերի վրա կատարված փորձերը բացահայտում են սեղմման գործակիցները և կատարման ժամանակները՝ ընդգծելով սեղմման մեթոդների կիրառման արդյունավետությունը, ինչպիսիք են Deflate-ը և Zstandard-ը: Նկար 4-ը ներկայացնում է Dask-ի և Spark-ի արտադրողականության համեմատությունը NDVI ինդեքսի հաշվարկի համար՝ հաշվի առնելով մուտքային տվյալների տարբեր ծավալները և սեղմման մեթոդները, որոնք աջակցվում են միջավայրերի կողմից:

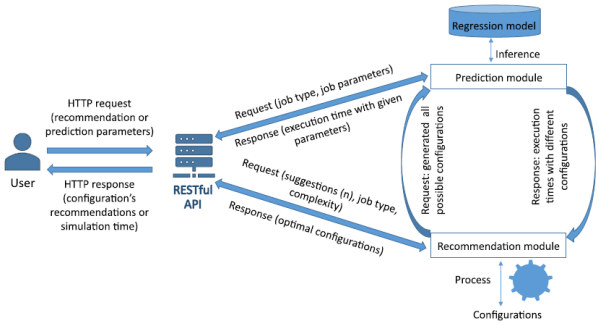


Նկար 4: Dask-ի և Spark-ի համեմատությունը՝ հաշվի առնելով 16, 32, 64 ԳԲ ծավալով մուտքային տվյալները և աջակցվող սեղմման մեթոդները

Կատարված գիտափորձերը ցույց են տալիս, որ Dask-ը և Spark-ը տրամադրում են տվյալների մշակման նմանատիպ կատարման ժամանակ, իսկ Dask և Zstandard սեղմման մեթոդների համադրումը տրամադրում է լավագույն արդյունքը, մասնավորապես՝ համադրումը նվազեցնում է օգտագործված պահոցի ծավալը 16%-ով և միաժամանակ արագացնում է կատարման ժամանակը 4,72 և 3,99 անգամ համապատասխանաբար Dask-ում և Spark-ում համեմատած Deflate մեթոդի հետ, որը սովորաբար լռելյայն օգտագործվում է գլոբալ ԵԴ տվյալների պահոցներում:

4.3 ենթագլուխը ներկայացնում է արտադրողականության օպտիմալացված որոշումների կայացման ծառայությունը: Ծառայությունը տրամադրում է տվյալների սեղմման մեթոդների ընտրության արդյունավետ առաջարկություններ՝ նպաստելով

պահեստավորման խնայողությանը և բարելավելով տվյալների մշակման արդյունավետությունը բաշխված հաշվարկների ընթացքում: Ծառայության ճարտարապետությունը ներկայացված է Նկար 5-ում:



Նկար 5: Արտադրողականության օպտիմալացված որոշումների կայացման ծառայության կառուցվածքը

Ծառայությունը բաղկացած է առաջարկությունների (Recommendation) և կանխատեսման (Prediction) մոդուլներից: Կանխատեսման մոդուլը գնահատում է տվյալների մշակման ժամանակը՝ հիմնվելով տարբեր գծային և բազմանդամ ռեգրեսիոն մոդելների վրա, որոնք ուսուցանվել են օգտագործելով սիմուլյացիոն տվյալների հավաքածուների մոդելավորման հիմնական հատկանիշները: Ուսուցանման ընթացքում նվազագույնի է հասցվել հետևյալ կորստի ֆունցիան.

$$L = (\ln(y) - X\beta)^T(\ln(y) - X\beta) + \lambda\beta^T\beta$$

Օգտագործելով գրադիենտային վայրէջքի մոտեցումը՝ կորստի ֆունկցիան նվազագույնի հասցնելու համար ուսուցման փուլում սահմանվել են β կշիռները: X-ը տարբեր հատկանիշներից բաղկացած տվյալների հավաքածուն է, y-ը տվյալ X հատկանիշների դեպքում կատարման ժամանակն է, իսկ λ-ն կանոնավորացման պարամետրն է: Բազմանդամի և կանոնավորացման հիպերպարամետրերի աստիճանը որոշվում է խաչաձև վավերացման (cross-validation) մեթոդի միջոցով: Ուսուցանումից հետո կանխատեսումը կատարվում է $y' = e^{X\beta}$բանաձևով, որտեղ y'-ը կատարման կանխատեսված ժամանակն է:

Առաջարկությունների մոդուլը հենվելով կանխատեսման մոդուլի տրամադրած արդյունքների վրա և հաշվի առնելով օգտագործողի հարցումը տրամադրում է արդյունավետ առաջարկություն: Մոդուլը դիտարկում է բոլոր հնարավոր կազմաձևերը՝ հաշվի առնելով տվյալների սեղմման աջակցվող մեթոդները տվյալ աշխատանքային հոսքի համար: Կանխատեսման մոդուլը կիրառվում է հնարավոր կազմաձևերի կատարման ժամանակները գնահատելու համար: Վերջապես,

առաջարկությունների մոդուլը դասավորում է ստացված արդյունքները ըստ կատարման ժամանակի և վերադարձնում ամենալավ n կազմաձևերը:

4.4 ենթագլուխը ամփոփում է 4-րդ գլխի արդյունքները:

Աշխատանքի հիմնական արդյունքները.

1. Մշակվել է առանց սերվերի, հաշվողական ենթակառուցվածքից անկախ և ընդլայնվող ԵԴ տվյալների մշակման համալիր համակարգ, որը բավարարում է մեծածավալ տվյալների արդյունավետ մշակման և պահպանման համար դիտարկված հիմնական կատարողական ցուցանիշներին [1, 4, 5]:
2. Մշակվել է ԵԴ տվյալների արդյունավետ մշակման համար բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ մեթոդ՝ հաշվի առնելով հաշվողական ենթակառուցվածքների առանձնահատկությունները և աշխատանքային հոսքերի բարդությունը [3]:
3. Մշակվել է ԵԴ տվյալների պահպանման համար նախատեսված արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն, որը տվյալների մշակման արտադրողականության բարձրացման համար առաջարկում է տվյալների սեղմման արդյունավետ մեթոդներ [2, 6, 7]:

Հրապարակված աշխատանքների ցանկ

1. H. Astsatryan, A. Lalayan, G. Giuliani, “Scalable data processing platform for earth observation data repositories”, Scalable Computing: Practice and Experience, 24(1), pp. 35-44, 2023. doi: 10.12694/scpe.v24i1.2041.
2. A. Lalayan, “Data Compression-Aware Performance Analysis of Dask and Spark for Earth Observation Data Processing”, Mathematical Problems of Computer Science, 59, pp. 35-44, 2023. doi: 10.51408/1963-0100.
3. A. Lalayan, H. Astsatryan, G. Giuliani, “A Multi-Objective Optimization Service for Enhancing Performance and Cost Efficiency in Earth Observation Data Processing Workflows”, Baltic Journal of Modern Computing, 11(3), pp. 420-434, 2023. doi: 10.22364/bjmc.2023.11.3.05.
4. H. Astsatryan, H. Grigoryan, R. Abrahamyan, A. Lalayan, S. Asmaryan, G. Giuliani, Y. Guiguz, “Scalable data processing and visualization service of Sentinel 5P for Earth Observations Data Cubes”, Arabian Journal of Geosciences, 16, 618, 2023. doi: 10.1007/s12517-023-11672-y.
5. A. Lalayan, H. Astsatryan, G. Giuliani, “Enhancing Earth Observation Data Processing through Optimized Multi-Modular Service”, Computer Science and Information Technologies (CSIT), pp. 95-98, 2023. doi: 10.51408/csit2023_19.
6. H. Astsatryan, A. Lalayan, A. Kocharyan, D. Hagimont, “Performance-efficient Recommendation and Prediction Service for Big Data frameworks focusing on Data Compression and In-memory Data Storage Indicators”, Scalable Computing: Practice and Experience, 22(4), pp. 401-412, 2021. doi: 10.12694/scpe.v22i4.1945.
7. H. Astsatryan, A. Kocharyan, D. Hagimont, A. Lalayan, “Performance optimization system for hadoop and spark frameworks”, Cybernetics and Information Technologies, 20(6), pp. 5-17, 2020. doi:10.2478/cait-2020-0056.

Разработка облачной и высокопроизводительной платформы для данных наблюдения Земли

Резюме

Данные наблюдения Земли (НЗ) представляют собой огромный объем информации, собранной со спутников, самолетов, дронов и наземных датчиков. Эти данные необходимы для мониторинга окружающей среды, предоставляя информацию о различных слоях Земли, включая атмосферу или водные ресурсы. Увеличение объема данных НЗ требует огромных вычислительных ресурсов для обработки крупномасштабных данных дистанционного зондирования.

Исследовательские сообщества либо используют специализированные платформы НЗ, или пользуются общими сервисами и вычислительными ресурсами глобальных инфраструктурных провайдеров, таких как Amazon, Google или Microsoft, которые предлагают удобные сервисы и огромные вычислительные мощности и ресурсы хранения данных. Оба подхода имеют свои преимущества и ограничения. Различные специализированные платформы предлагают комплексные решения для доступа, обработки и визуализации данных НЗ, такие как Sentinel Hub (SH), Google Earth Engine (GEE), WEkEO, CREODIAS и др. Эти платформы интегрированы с вычислительными инфраструктурами, предлагаемыми мировыми облачными провайдерами. В частности, SH работает на базе Amazon Web Services, GEE размещается на Google Cloud Platform, а облачные платформы WEkEO и CREODIAS, созданные по инициативе Copernicus, - на CloudFerro. Таким образом, подобные решения являются общими и доступны только пользователям тех инфраструктур, которые платят за используемые ресурсы. Существенным ограничением является также привязка к вендорам, поскольку переход на другую платформу или облачного провайдера из-за наличия проприетарных форматов и инструментов затруднен и требует больших затрат.

Чтобы уменьшить зависимость от поставщиков и общих решений, обычно используется реализация концепции куба данных наблюдения с открытым исходным кодом, известная как платформа Open Data Cube (ODC). ODC организует данные в формате кубов, что облегчает проведение пространственно-временного анализа. Являясь самостоятельной платформой и не завися от конкретных вычислительных инфраструктур, ODC обеспечивает гибкость, позволяя пользователям создавать свои экземпляры и адаптировать систему к своим уникальным потребностям. Однако среда ODC сталкивается с проблемами горизонтальной масштабируемости, производительности и другими факторами, необходимыми для эффективной работы с растущими потребностей в данных НЗ. Кроме того, затруднено автоматическое предоставление сторонних вычислительных инфраструктур с ODC.

Учитывая указанные ограничения, необходимо разработать независимую от вычислительной инфраструктуры комплексную систему обработки данных наблюдения Земли, обеспечивающую гибкие и масштабируемые решения с учетом важнейших ключевых показателей эффективности.

Цель и рассматриваемые задачи

Основной целью работы является разработка масштабируемой комплексной системы обработки данных НЗ, не зависящей от вычислительных инфраструктур. Для достижения этой цели рассматриваются следующие задачи:

1. Разработать масштабируемую и бессерверную комплексную систему обработки данных НЗ, взаимодействующую с хранилищами данных и облачными высокопроизводительными инфраструктурами и не зависящую от вычислительных инфраструктур, обеспечивающую эффективную и гибкую обработку данных.
2. Разработать метод многообъектной оптимизации при выборе распределенного вычислительного кластера для обработки данных НЗ, обеспечивающий масштабирование вычислительных ресурсов по требованию с учетом различных факторов, таких как производительность и стоимость.
3. Оценить влияние методов сжатия данных на распределенную обработку больших данных НЗ, обеспечивая баланс между экономией на хранилище и улучшением скорости обработки.

Практическая значимость работы

Разработанная комплексная система может быть использована для эффективной обработки крупномасштабных данных НЗ с учетом производительности обработки данных и факторов стоимости с использованием облачных или высокопроизводительных инфраструктур.

Структура и объем работы

Диссертация состоит из введения, 4 глав и списка использованной литературы. Диссертация написана на 109 страницах и имеет 124 ссылки на литературу.

Основные результаты работы

1. Разработана масштабируемая, бессерверная и независимая от вычислительной инфраструктуры комплексная система обработки данных НЗ, удовлетворяющая ключевым показателям производительности для эффективной обработки и хранения крупномасштабных данных [1, 4, 5].
2. Разработан многоцелевой метод выбора распределенного вычислительного кластера для эффективной обработки данных НЗ с учетом особенностей вычислительных инфраструктур и сложности рабочих процессов [3].
3. Разработан оптимизированный по производительности сервис принятия решений для хранения данных НЗ, рекомендующий эффективные методы сжатия данных для повышения производительности их обработки [2, 5, 6, 7].

Development of a cloud and high-performance platform for earth observation data

Abstract

Earth observation (EO) data represent the vast amount of information gathered from satellites, airplanes, drones, and ground-based sensors. This data is essential for aiding environmental monitoring, providing information on different layers of the Earth, including the atmosphere or water resources. Increasing EO data requires enormous computing resources to process large-scale remote sensing data.

Research communities either utilize specialized EO platforms or leverage shared services and computing resources from global infrastructure providers like Amazon, Google, or Microsoft, which offer user-friendly services along with enormous computational and storage capacities. Both approaches have their advantages and limitations. Various specialized platforms offer comprehensive solutions for accessing, processing, and visualization of EO data, such as Sentinel Hub (SH), Google Earth Engine (GEE), WEkEO, CREODIAS, etc. These platforms are integrated with the computing infrastructures offered by global cloud providers. Specifically, SH operates on Amazon Web Services, GEE is hosted on the Google Cloud Platform, and the cloud platforms WEkEO and CREODIAS, initiated by Copernicus, are situated on CloudFerro. Therefore, such solutions are general and available only to users of those infrastructures that pay for the used resources. Vendor lock-in is also a significant limitation, as switching to another platform or cloud provider due to proprietary formats and tools is difficult and expensive.

To reduce reliance on vendors and generic solutions, the common approach is to employ an open-source implementation of the Earth Observation Data Cube (EODC) concept, known as the Open Data Cube (ODC) platform. ODC organizes data in cube format, facilitating spatial and temporal analyses. As a standalone platform and independent of the specific computing infrastructures, ODC offers flexibility, allowing users to establish their instances and customize the system to their unique needs. However, the ODC environment faces challenges in terms of horizontal scalability, performance, and other factors crucial for efficiently handling the expanding needs of EO data. Besides that, the automatic provision of third-party computational infrastructures with ODC is complicated.

Considering the mentioned limitations, it is necessary to develop a complex EO data processing system independent of the computing infrastructure, which provides flexible and scalable solutions, taking into account crucial key performance indicators.

The purpose and problems of the work

The main purpose of the work is to develop a scalable EO data processing complex system independent of computing infrastructures. To achieve this goal, we consider the following problems:

1. Develop a scalable and serverless EO data processing complex system that is interoperable with data repositories and cloud-HPC infrastructures and is independent of computing infrastructure, ensuring efficient and flexible data processing.

2. Develop a multi-objective optimization method for choosing a distributed computing cluster for EO data processing, which will provide on-demand scaling of computing resources, taking into account various factors such as performance and cost.
3. Evaluate the impact of data compression techniques on distributed Big EO Data processing, striking a balance between storage savings and processing speed improvements.

The practical significance of the work

The developed complex system can be used for efficient processing of large-scale EO data taking into account data processing performance and cost factors using cloud or HPC infrastructures.

Structure and scope of work

The dissertation consists of an introduction, 4 chapters, and a list of used literature. The thesis is written in 109 pages and has 124 literature references.

The main results of the work

1. A scalable, serverless, and computing infrastructure-independent EO data processing complex system is developed to meet key performance indicators for efficiently processing and storing large-scale data. [1, 4, 5].
2. A multi-objective method is developed for selecting a distributed computing cluster for efficient processing of EO data, considering the specific characteristics of computing infrastructures and the complexity of workflows [3].
3. A performance-optimized decision-making service for EO data storage is developed, recommending effective data compression methods to enhance data processing performance [2, 5, 6, 7].