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 (ՀԱՊՀ)

Գիտական ղեկավար՝ տեխ. գիտ. դոկտոր Հ.Վ. Ասցատրյան
Պաշտոնական ընդդիմախոսներ՝ տեխ. գիտ. դոկտոր X X X
տեխ. գիտ. դոկտոր X X X
Առաջատար կազմակերպություն՝ XXX

Ատենախոսության պաշտպանությունը տեղի կունենա 2023թ. X X-ին ժամը X: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 X X-ին:

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

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

Scientific supervisor: H. Astsatryan, PhD, Doctor of Sciences
Official opponents: XXX
XXX

Leading organization: XXX

The Defense will take place on X X, 2023; at XX: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 X X, 2023.

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

Աշխատանքի ընդհանուր նկարագիրը

Թեմայի արդիականությունը. Երկրի դիտարկման (ԵԴ) տվյալները ներկայացնում են արբանյակներից, ինքնաթիռներից, անօդաչու թռչող սարքերից և ցամաքային տվիչներից հավաքված տեղեկատվության հսկայական քանակություն[1]: Այս տվյալները կարևոր են շրջակա միջավայրի մշտադիտարկման համար, քանի որ այն տրամադրում է տեղեկատվություն Երկրի աշխարհագրական թաղանթների մասին, ինչպիսիք են մթնոլորտը կամ ջրային ռեսուրսները[2]: ԵԴ տվյալները տրամադրում են ժամանակային շարքի տվյալներ, որոնք կարևոր են դիտարկելու շրջակա միջավայրի երկարաժամկետ փոփոխությունները, կլիմայի փոփոխության և մարդու գործունեության հետևանքները: Տարբեր տվիչների ինտեգրումը հնարավորություն է տալիս հետազոտել Երկրի գործընթացների լայն տիրույթ՝ սկսած ցամաքի ծածկույթի փոփոխություններից մինչև մթնոլորտային երևույթներ: ԵԴ տվյալների բարդությունն ու ծավալը աճում են՝ ստեղծելով դժվարություններ պահպանման, կառավարման և մշակման մակարդակներում:

ԵԴ տվյալների ծավալի աճը պահանջում է մեծ ծավալի հաշվողական ռեսուրսներ: Հետազոտական համայնքները և օգտագործողները առաջարկում են տեղական բարձր արտադրողականությամբ հաշվողական (high-performance computing - HPC) և ամպային ենթակառուցվածքներ կամ օգտագործում են գլոբալ ամպային մատակարարների ծառայությունները՝ ԵԴ տվյալների մշակման աճող կարիքները բավարարելու համար, որը գնալով դառնում է ավելի բարդ և պահանջում է ավելի շատ ապարատային ռեսուրսներ և ճկուն ծրագրակազմ: Առկա է ԵԴ տվյալների կառավարման երկու մոտեցումներ: Մի կողմից կարելի է դիտարկել գլոբալ ռեսուրսների մատակարարների կողմից առաջարկվող ընդհանուր ծառայությունների օգտագործումը: Մյուս կողմից ենթադրում է մասնագիտացված հարթակների տեղակայումը: Երկու մոտեցումներն էլ ունեն իրենց առավելություններն ու սահմանափակումները: Թեև առաջարկելով օգտագործողներին հարմար ծառայություններ, ինչպես նաև հաշվողական և պահեստավորման հսկայական հնարավորություններ, գլոբալ ամպային մատակարարները, ինչպիսիք են Amazon-ը, Google-ը կամ Microsoft-ը, տալիս են ընդհանուր լուծումներ և պահանջում են վճարել տվյալների պահպանման և մշակման ընթացքում ռեսուրսների օգտագործման համար: Մատակարարի արգելափակման խնդիրը (vendor lock-in) ևս զգալի սահմանափակում է, քանի որ այլ ամպային մատակարարի անցնելը սեփական ձևաչափերի և գործիքների պատճառով կարող է դժվար և թանկ լինել:

Տարբեր ԵԴ հարթակներ, ինչպիսիք են Sentinel Hub-ը[3], Open Data Cube-ը (ODC)[4], OpenEO-ն[5] և այլն, առաջարկում են համապարփակ լուծումներ ԵԴ

[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] Sentinel Hub-ի կայքէջն է. https://www.sentinel-hub.com/
[4] Open Data Cube-ի դոկումենտացիան հասանելի է. https://www.opendatacube.org/.

տվյալների հասանելիության, մշակելու և վիզուալիզացիայի համար: Հարկ է նշել, որ ODC-ն միակ բաց կոդով հարթակն է, և այս փակ լուծումների հետևանքով, հարմարեցման և ընդլայնման տարբերակները սահմանափակվում են՝ ի վերջո սահմանափակելով հարթակների օգտագործողներին ընդհանուր լուծումներով և նվազեցնելով նրանց ընդհանուր ճկունությունը: Ավելին, ինչպես գլոբալ ամպային մատակարարները, այս լուծումները նույնպես ներառում են վճարներ, որոնք օգտվողները պետք է վճարեն ծառայության հասանելիության համար, և մատակարարի արգելափակման խնդիրը այս դեպքում նույնպես մարտահրավեր է:

ODC-ն ԵԴ տվյալների խորանարդ (EO Data Cube)[6] հայեցակարգի բաց կոդով իրականացում է, նախատեսված մեծածավալ ԵԴ տվյալներ դժվարությունները հաղթահարելու համար, և տրամադրում է տվյալների վերլուծության, ինդեքսավորման, որոնման, ընդունման, պահեստավորման, մշակման և արտացոլման համար ճկուն ծրագրակազմ: ODC-ն ծառայում է որպես տվյալների պահոց և տիրույթին հատուկ ծառայությունների մատակարար, գործում է անկախ որպես ինքնուրույն հարթակ և կախված չէ հաշվողական ենթակառուցվածքներից: Արդյունքում, հիմնական նկատառումները, ինչպիսիք են ընդլայնելիությունը, արտադրողականությունը և այլ գործոններ հաշվի չեն առնվում, քանի որ այն հորիզոնական ընդլայնվող չէ և կարող է տեղակայվել մեկ սերվերի վրա, իսկ երրորդ կողմի (third-party) հաշվողական ենթակառուցվածքների ավտոմատ տրամադրումը ODC-ին բարդ է: Հետևաբար, նշված սահմանափակումների հաղթահարումը փաստացի մարտահրավեր է:

Աշխատանքի նպատակն է մշակել ԵԴ տվյալների մշակման համալիր համակարգ, որը ԵԴ տվյալների արդյունավետ մշակման համար հաղթահարում է նշված սահմանափակումները և ապահովում է ճկուն և ընդարձակելի լուծումներ՝ հաշվի առնելով արտադրողականության կարևորագույն հիմնական ցուցանիշները:

Աշխատանքի նպատակը և դիտարկված խնդիրները: Աշխատանքի հիմնական նպատակն է մշակել ԵԴ տվյալների մշակման ընդլայնվող և օպտիմիզացված 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 ԵԴ տվյալների մշակման համակարգ, որն անխափան կերպով միավորում է տվյալների պահոցները ամպային և բարձր արտադրողականությամբ ենթակառուցվածքներին՝ ապահովելով ԵԴ տվյալների արդյունավետ և ճկուն մշակումը:
2. Մշակել բաշխված հաշվողական կլաստերի ընտրության բազմաֆունկցիոնալ օպտիմալացման մեթոդ՝ հիմնված տարբեր չափանիշների վրա, ներառյալ արտադրողականություն և ծախսեր:

[5] Open EO-ի կայքէջն է. https://openeo.org/.
[6] 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.

3. Գնահատել բաշխված մեծածավալ ԵԴ տվյալների մշակման վրա տվյալների սեղմման մեթոդների ազդեցությունը՝ հավասարակշռություն հաստատելով պահեստավորման խնայողության և մշակման արագության բարելավման միջև:

Հետազոտության օբյեկտները: Այս աշխատությունում հետազոտության հիմնական օբյեկտներն են.

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

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

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

1. ԵԴ տվյալների մշակման համալիր համակարգ, որը հաշվի առնելով միջազգային հիմնօրինակները ճկուն կերպով միավորում է ԵԴ տվյալների պահոցները ընդլայնելի հաշվողական ռեսուրսների հետ:
2. Ամպային և բարձր արտադրողականությամբ հաշվողական համակարգերի արտադրողականության, ծախսարդյունավետության և այլ գործոնների գնահատման համար բազմաֆունկցիոնալ օպտիմալացման մեթոդ, որը հաշվի է առնում ենթակառուցվածքների առանձնահատկությունները և աշխատանքային հոսքերի բարդությունը:
3. Արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն, որն առաջարկում է տվյալների սեղմման արդյունավետ մեթոդներ միաժամանակ հաշվի առնելով պահեստային տարածքի խնայողությունը և ԵԴ տվյալների մշակման արտադրողականությունը:

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 ընդլայնվող համակարգ, որը միավորում է ԵԴ տվյալների պահոցները ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների հետ:
2. Բաշխված հաշվողական կլաստերի ընտրության համար բազմաֆունկցիոնալ օպտիմալացման մեթոդ՝ տվյալների մշակման արտադրողականության և ծախսարդյունավետության լավարկման համար:
3. Արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն՝ սեղմման արդյունավետ մեթոդներ առաջարկելու համար, հաշվի առնելով պահեստային տարածքը և ԵԴ տվյալների մշակման արդյունավետությունը:

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 ԵԴ տվյալների ներածությունը և նշանակությունը: Գլուխը ներառում է նաև հաշվողական ենթակառուցվածքները,

հարթակները և ծառայությունները, հինօրինակները, գործիքները և տվյալների ձևաչափերը, որոնք անհրաժեշտ են լայնածավալ ԵԴ տվյալների արդյունավետ կառավարման և մշակման համար:

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

1.2 ենթագլխում նկարագրված է շրջակա միջավայրի մշտադիտարկման համար ԵԴ տվյալների նշանակությունը: Այս ենթագլուխն ուսումնասիրում է ԵԴ տվյալների հետ աշխատանքին բնորոշ առավելություններն ու մարտահրավերները, որը հնարավորություն է տալիս պատկերացում կազմել բաց տվյալներ տրամադրող արբանյակների մասին՝ պարզաբանելով դրանց բնութագրերը, տվիչային գոտիները և ԵԴ ինդեքսները, որոնք ըստ էության հաշվարկներ կամ ալգորիթմներ են, ստացվում են հեռահար զոնդավորման տեղեկատվությունից և նախատեսված են բնապահպանական պայմանների մասին տեղեկատվություն ստանալու համար:

1.3 ենթագլխում ներկայացված են առանց սերվերի, ամպային և բարձր արտադրողականությամբ հաշվողական ենթակառուցվածքները: Այս ենթագլուխը ներկայացնում է ամպային լուծումներ, որոնք արդյունավետորեն կիրառվում են ԵԴ տվյալները կառավարելու համար: Նաև ներկայացված է բաշխված մշակման հայեցակարգը՝ ընդգծելով բաշխված և զուգահեռ մեծածավալ տվյալների հավաքածուները մշակելու համար լայնորեն կիրառվող միջավայրերը, ինչպիսիք են Apache Hadoop-ը[7], Spark-ը[8] և Dask-ը[9]:

Այս ենթագլուխը նկարագրում է նաև ԵԴ գլոբալ պահոցները, ինչպես նաև գլոբալ ամպային մատակարարների և լայնորեն օգտագործվող ԵԴ տվյալների հարթակների և ծառայությունների կողմից տրամադրված լուծումները: Ենթագլուխը գնահատում է լուծումների առավելությունները և թերությունները, ինչպես նաև ուսումնասիրվում է բաց կոդով ODC ծրագրային ապահովման հնարավորությունները, որը պարզեցնում է մեծածավալ և բարդ ԵԴ տվյալների մշակումն ու վերլուծությունը: Հատկանշական է, որ ԵԴ տվյալների կառավարման և մշակման համար այս հարթակը ներդրվել և օգտագործվել է տարբեր երկրներում, այդ թվում՝ Ավստրալիայում, Շվեյցարիայում, Բրազիլիայում և Հայաստանում: Ներկայացված են ODC-ի կիրառման որոշ օրինակներ՝ ցուցադրելով ODC-ի կիրառման արդյունավետությունը բնապահպանական տարբեր մարտահրավերներին դիմակայելու և գլոբալ մշտադիտարկման ջանքերը հեշտացնելու համար:

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

[7] Apache Hadoop-ի կայքէջն է. https://hadoop.apache.org/.
[8] Apache Spark-ի կայքէջն է. https://spark.apache.org/.
[9] Dask-ի կայքէջն է. https://www.dask.org/.

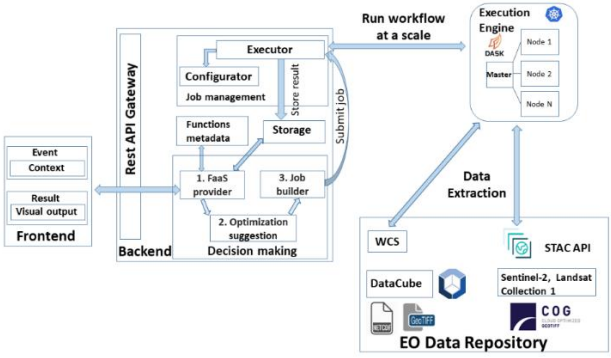
գործիքներին, որոնք առանցքային դեր են խաղում տվյալների մշակման, պահպանման և տարածման գործում:

1.5 ենթագլուխը եզրափակում է առաջին 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 մշակված ընդլայնվող ԵԴ տվյալների մշակման համալիր համակարգը[10]:

2.1 ենթագլուխը ներկայացնում է ԵԴ տվյալների արդյունավետ մշակման համար հիմնական կատարողական ցուցանիշները, ինչպիսիք են արտադրողականությունը և ընդլայնելիությունը, ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների և ԵԴ տվյալների պահոցների փոխգործունակությունը, ավտոմատ և արագ ամպային և բարձր արտադրողականությամբ ռեսուրսների տրամադրումը և ընդլայնումը, ծախսերի արդյունավետություն, նոր ձևաչափերի և լուծումների աջակցումը և կապը ODC-ի հետ: Ներկայացված է այլ հեղինակների կողմից առաջարկվող լուծումների համապարփակ ուսումնասիրություն, ինչպես նաև այս լուծումների սահմանափակումների գնահատումը սահմանված հիմնական կատարողական ցուցանիշների համատեքստում:

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



Նկար 1: ԵԴ տվյալների մշակման ընդլայնվող համալիր համակարգի կառուցվածքը

Համակարգը բաղկացած է 4 հիմնական մոդուլներից՝ Frontend, Backend, Execution engine և EO data repositories: Frontend մոդուլը օգտագործողի միջերեսն է,

10 Կոդը հասանելի է https://github.com/ArmHPC/Scalable-EO-system

որն ապահովում է Backend-ի հետ փոխազդեցությունը և արդյունքների վիզուալիզացիան: Օգտագործողների հարցումները մշակելու համար Backend-ը տրամադրում է RESTful API, հաշվի առնելով տարածքը, ժամանակահատվածը և մշակման ֆունկցիայի պարամետրերը: Execution engine մոդուլը ապահովում է ԵԴ տվյալների մշակման ընդլայնելիությունը տվյալների բաշխված մշակման միջոցով՝ օգտագործելով բարձր արտադրողականությամբ հաշվողական կլաստերներ, որոնք հիմնված են ամպային կամ բարձր արտադրողականությամբ ռեսուրսների վրա: EO data repositories մոդուլը տվյալների պահոցներն են, որոնք պահոցների փոխգործունակությունն ապահովելու համար տրամադրում են ԵԴ տվյալների հասանելիություն, ինչը համակարգին դարձնում է ավելի ճկուն, ընձեռնելով հնարավորություն աշխատել տարբեր պահոցների հետ: Համակարգի արդյունավետությունը ցուցադրվում է դիտարկելով նորմալացված տարբերության բուսականության ինդեքսը[11] (Normalized Difference Vegetation Index - NDVI), որը բուսականության խտությունը որոշելու գրաֆիկական ցուցիչ է:

2.3 ենթագլուխը ներկայացնում է համակարգի տրամադրումը որպես ծառայություն օդի աղտոտվածության մշտադիտարկման նպատակով, որը նախատեսված է ODC-ների համար:

2.4 ենթագլխում համառոտ ամփոփվում են 2-րդ գլխում ստացված արդյունքները:

Երրորդ գլուխը ներկայացնում է ԵԴ տվյալների մշակման աշխատանքային հոսքերի համար բազմաֆունկցիոնալ օպտիմալացման մեթոդ, որը հնարավորություն է տալիս ընտրել տվյալների բաշխված մշակման արդյունավետ կլաստեր, հաշվի առնելով արտադրողականությունը և ծախսարդյունավետությունը:

3.1 ենթագլխում ներկայացվում է խնդրի ձևակերպումը՝ ընդգծելով բազմաֆունկցիոնալ օպտիմալացման նշանակությունը ԵԴ տվյալների մշակման աշխատանքային հոսքերում: ԵԴ տվյալների մշակման առաջադրանքների կատարման ժամանակը կախված է կլաստերի կազմաձևից, մասնավորապես աշխատող հանգույցների քանակից և հաշվողական բնութագրերից (օրինակ՝ միջուկների քանակ և օպերատիվ հիշողություն - ՕՀ): Միևնույն ժամանակ ամպային մատակարարները առաջարկում են տարբեր տեսակի հաշվողական ռեսուրսների օրինակներ, որոնք ունեն տարբեր քանակի միջուկներ և ՕՀ, որոնցից յուրաքանչյուրի կիրառումը ենթադրում է տարբեր չափի ծախսեր: Dask-ի օպտիմալ կազմաձև ընտրելիս արտադրողականության և արժեքի միջև օպտիմալ փոխզիջում գտնելը դժվար է: Մի կողմից ավելի շատ հաշվողական ռեսուրսների օգտագործումը կարող է հանգեցնել տվյալների մշակման ավելի մեծ զուգահեռացման և ժամանակի կրճատմանը, իսկ մյուս կողմից այն ավելի մեծ ծախսեր է առաջացնում ամպային մատակարարների կողմից: Բացի այդ, հնարավոր կլաստերի

[11] 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*, vol. 10, 197, 2023.

կազմաձևերի բազմությունը տեսականորեն անվերջ է, հետևաբար օպտիմալ կազմաձև ընտրելը NP դասի խնդիր է:

ԵԴ տվյալների մշակման առաջադրանքի արտադրողականության և ծախսերի նպատակները կարող են ներկայացվել հետևյալ բանաձևերով.

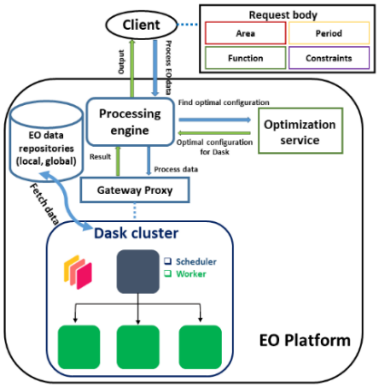
$$t = \tau(s, n, r)$$
$$p = \upsilon(t, n, r)$$
$$r \in R;\ n, s \in N$$

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: Հաշվողական Dask կլաստերը բաղկացած է n թվով հանգույցներից, որոնցից յուրաքանչյուրն ունի r տեսակի հաշվողական օրինակներ դիտարկված R վերջավոր շարքից: s-ը և n -ը N բնական թվերի բազմությունից են: ԵԴ տվյալների մշակման առաջադրանքի արժեքը՝ p-ն կախված է առաջադրանքի կատարման ժամանակից և հաշվողական կլաստերի բնութագրերից, մասնավորապես հանգույցների քանակից և հանգույցների օրինակի տեսակից: Ստորև բերված բանաձևն օգտագործվում է հաշվողական ռեսուրսների օպտիմալ համակցությունը գտնելու համար՝ հաշվի առնելով արտադրողականության և ծախսերի նպատակները:

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

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

3.2 ենթագլուխը ներկայացնում է ԵԴ տվյալների մշակման համալիր համակարգի բազմաֆունկցիոնալ օպտիմալացման մեթոդը (տես Նկար 2):



Նկար 2: ԵԴ տվյալների մշակման համակարգի կառուցվածքը

Օպտիմալացման մեթոդը հնարավորություն է տալիս ստեղծել օպտիմալ կլաստերային կազմաձևեր ԵԴ տվյալների մշակման արտադրողականության և ծախսարդյունավետության լավարկման համար:

Օպտիմալացման մեթոդի ինտեգրումը ԵԴ տվյալների մշակման համակարգում տրամադրում է հետևյալ աշխատանքային հոսքը.

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

Օպտիմալացման մեթոդը հիմնված է բազմաֆունկցիոնալ Պարետո օպտիմալացման մեթոդի[12] վրա, որը 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: Օպտիմալացման մեթոդի ալգորիթմական տեսքը ներկայացված է ստորև:

[12] S. Petchrompo, D. W. Coit, et al. A review of Pareto pruning methods for multi-objective optimization, Computers & Industrial Engineering, vol. 167, 108022, 2022.

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 ← finite set of Dask cluster configurations
results ← {}
optimalPoints ← {}
for config in configs do
    time ← τ(s, n, r) ▷ find or predict execution time for the given config and complexity s
    cost ← time × config.instanceRate × config.nodes
    if cost ≤ p' AND time ≤ t' then
        results.append((config, cost, time))
    end if
end for
for r in results do
    nonDominatedPoints ← {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
```

Օպտիմալացման մեթոդը սկսվում է առաջադրանքի կատարման ժամանակի և պահանջվող հաշվողական ռեսուրսների արժեքի գնահատմամբ՝ հաշվի առնելով Dask կլաստերի տարբեր կազմաձևեր դիտարկվող վերջավոր հավաքածուից: 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 են, հաշվի առնելով մուտքային տվյալների ծավալը: 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 օրինակի ժամային դրույքաչափով: Ենթադրենք, որ ստացված արժեքը և կատարման ժամանակը համընկնում են օգտագործողի սահմանափակումների հետ (t' - կատարման ժամանակի սահմանափակում և p' - ծախսերի սահմանափակում): Այդ դեպքում այն ավելացվում է օգտագործողի հարցմանը բավարարող լուծումների ցանկում: Կազմաձևերի ստացված ցանկը զտվում է՝ վերացնելով այն տարրերը, որոնք գերազանցում են մյուսներին՝ միաժամանակ հաշվի առնելով ծախսերի և արտադրողականության նպատակները: Մասնավորապես, ավելի բարձր արժեքով և կատարման ժամանակով կազմաձևերը հանվում են ցանկից: Այս գործողությունն

ավարտվելուն պես, չգերակշռող կազմաձևերի արդյունքում կազմված ցանկը ներառում է այնպիսի կազմաձևեր, որոնք առաջարկում են համեմատելի արժեք և արդյունավետություն, ինչպես նաև չունեն էական առավելություններ միմյանց նկատմամբ:

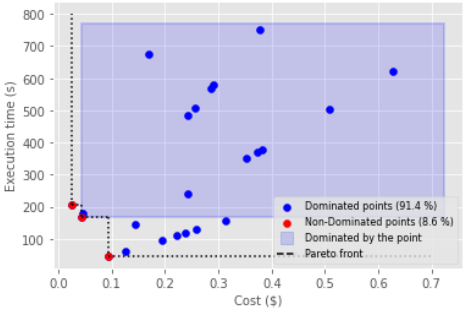
Արդյունավետությունը բարձրացնելու և ծախսերը կրճատելու համար անհրաժեշտ է կատարել զգալի թվով հաշվարկներ և փորձեր: Մոդելավորումների անցկացումը, որոնք ներառում են տատանվող բեռներ, արտադրողականություն կամ համակարգի չափսեր, կարող են թանկ լինել, և սիմուլյատոր գործիքները լուծում են տալիս այս ծախսը մեղմելու համար: Ենթագլուխը ներկայացնում է EO Cloud Simulator (EOCSim) նոր մոդել՝ ԵԴ տվյալների մշակման աշխատանքային հոսքերի համար, որը հիմնված CloudSim[13] գործիքի վրա, հանդիսանալով ընդլայնվող բազմամոդուլային ԵԴ համալիր համակարգի կարևոր մաս: CloudSim-ը գործում է ամպային հաշվողական միջավայրում՝ ամպային բարդ համակարգերի մոդելավորման և վերլուծության համար տրամադրելով մոդելավորման միջավայր: Այն հիմնված է միլիոն հրահանգներ վայրկյանում (Million instructions per second-MIPS) հատկության վրա, և հաշվի է առնում այլ գործոններ, ինչպիսիք են ընդհանուր ծախսերը, քաղաքականության իրականացումը, պլանավորման ռազմավարությունները և ամպային միջավայրում այլ կարևոր ասպեկտներ մոդելավորված: Սա հնարավորություն է տալիս հետազոտողներին և մշակողներին համակողմանիորեն գնահատել և օպտիմալացնել ամպի վրա հիմնված լուծումների կատարումն ու արդյունավետությունը: Մոդելն ցուցադրում է իր արդյունավետությունը՝ առաջարկելով գնահատումներ կատարման ժամանակի և ծախսերի համար, որոնք ներգրավված են նկարագրված կլաստերում ԵԴ տվյալների բաշխված մշակման մեջ:

3.3 ենթագլուխը գնահատում է առաջարկվող մեթոդը և քննարկում գնահատման արդյունքները: Գնահատումները կատարվում են օգտագործելով ինչպես CloudLab-ի[14], այնպես էլ հայկական ամպային ենթակառուցվածքի[15] հաշվողական ռեսուրսները, որոնք առաջարկում են ծառայությունների տարատեսակ համայնքների: Առաջարկվող օպտիմալացման մեթոդը ենթարկվել է գնահատման բազմաթիվ Dask կլաստերների վրա՝ ընդգրկելով աշխատող հանգույցների և օրինակների բազմազան տեսակները: Օգտագործվել են մեկ միջուկից և 2 ԳԲ օպերատիվ հիշողություն ունեցող հաշվողական օրինակից մինչև 64 միջուկ և 128 ԳԲ օպերատիվ հիշողությամբ օրինակ, և նույն Dask կազմաձևերը կիրառվել է բոլոր հնարավոր օրինակների վրա: Օպտիմալացման ծառայությունը գնահատելու համար մենք դիտարկում ենք երեք աշխատանքային

[13] 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), Pune, India, pp. 263-268, 2020.
[14] D. Duplyakin, R. Ricci, A. Maricq, et al. The Design and Operation of CloudLab. In Proceedings Of The USENIX Annual Technical Conference (ATC), 2019.
[15] 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.

ծանրաբեռնվածություն՝ տարբեր մուտքային տվյալների ծավալներով, որոնցից յուրաքանչյուրը համապատասխանում է Հայաստանի տարածքի տարբեր ժամանակաշրջանին: Ընտրված ծանրաբեռնվածությունն ունի շաբաթական (թեթև - 0,08 ՏԲ), ամսական (միջին - 0,32 ՏԲ) և սեզոնային (ծանր 1,2 ՏԲ) ծանրաբեռնվածություն:

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:

Օպտիմալացման ալգորիթմի արդյունքներն ամփոփելիս պարզվեց, որ յուրաքանչյուր ծանրաբեռնվածության համար կազմաձևորի միայն չնչին տոկոսն է համարվում Պարետո-օպտիմալ և ոչ գերակշռող: Մասնավորապես, սեզոնային ծանրաբեռնվածության համար կազմաձևերի միայն 5,7%-ն է Պարետո-օպտիմալ (երկու օպտիմալ կետեր՝ մեկը ապահովում է լավագույն կատարման ժամանակ, իսկ մյուսը՝ նվազագույն արժեք): Պարետո-օպտիմալ լուծումների քանակը կախված է ուսումնասիրված Dask կազմաձևերի բարդությունից: Ի հակադրություն, շաբաթական և ամսական ծանրաբեռնվածության առկա տարբերակներից հայտնաբերվեցին միայն երեք Պարետո-օպտիմալ կազմաձևեր: Այս գնահատումները ցույց են տալիս, որ օպտիմալացման ալգորիթմը արդյունավետորեն հայտնաբերում է արդյունավետ Dask կլաստերի կազմաձևեր յուրաքանչյուր ծանրաբեռնվածության համար: Այն կարող է օգնել օգտագործողներին որոշելու կազմաձևը՝ ելնելով արտադրողականության և ծախսերի նպատակներից: Օպտիմալ կազմաձևերից որևէ մեկի ընտրությունը կարող է բարելավել արտադրողականությունը՝ միաժամանակ նվազեցնելով հաշվողական ռեսուրսների արժեքը: Սեզոնային ծանրաբեռնվածությունը ուսումնասիրելիս, բոլոր դիտարկված կլաստերի կազմաձևերի համար կատարման միջին ժամանակը և արժեքը համապատասխանաբար կազմում են 201 վայրկյան և 0,405 դոլար: Ի հակադրություն, օպտիմալ կազմաձևերը ապահովում են 121 վայրկյան միջին կատարման ժամանակ և 0,17 դոլար արժեքը: Համեմատությունը

ցույց է տալիս, թե ինչպես օպտիմալ կազմաձև ընտրելը կարող է լավարկել արտադրողականությունը գրեթե 1,66 անգամ, մինչդեռ ծախսերը կրճատել միջինը 2,38 գործակցով: Շաբաթական ծանրաբեռնվածության դեպքում հնարավոր է հասնել միջինը 2,3 անգամ ավելի արագ կատարման ժամանակի և 4,7 անգամ կրճատված ծախսերի:

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

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

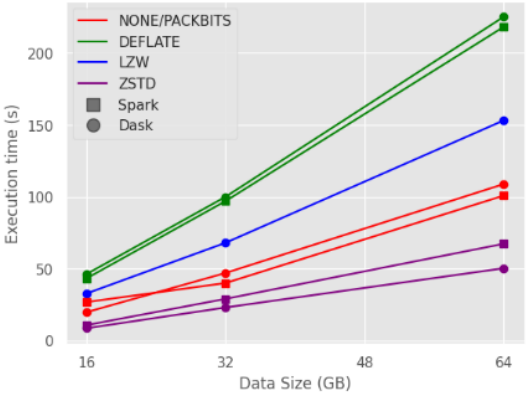
Գլուխ 4-ում ներկայացվում է լայնածավալ ԵԴ տվյալների հավաքածուների համար արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն: Այն հնարավորություն է տալիս ընտրել տվյալների սեղմման արդյունավետ մեթոդ, ինչի արդյունքում հնարավոր է խնայել պահեստավորման տարածքը և միևնույն ժամանակ բարելավել մշակման արտադրողականությունը:

4.1 ենթագլխում ներկայացված է աշխատանքի նախապատմությունը, ներառյալ լայնածավալ տվյալների մշակման օպտիմալացման գոյություն ունեցող մոտեցումների և լուծումների համառոտ ակնարկը: Տվյալների սեղմման մեթոդները օգտագործվում է տվյալների պահպանման և ցանցի թողունակության սահմանափակումները հաղթահարելու համար, երբ խոսքը գնում է մեծածավալ տվ յալների մշակման մասին: Մեծածավալ տվյալների ենթակառուցվածքներում այն նվազեցնում է տվյալների բլոկների ծավալը, որպեսզի նվազագույնի հասցնի մուտք/ելքի գործողության պատճառով պարտադրված ժամանակի հետաձգումը և խնայելու տարածք սկավառակների վրա: Խնդիրը օպտիմալ փոխզիջում գտնելն է, քանի որ սեղմման բարձր գործակիցը կարող է թերբեռնել մուտք/ելքը, բայց ծանրաբեռնել պրոցեսորը, մինչդեռ սեղմման թույլ գործակիցը կարող է թերբեռնել պրոցեսորը, բայց ծանրաբեռնել մուտք/ելքը:

4.2 ենթագլխում ներկայացված են գնահատումներ, որոնք ընդգծում են տվյալների սեղմման արդյունավետությունը մեծածավալ ԵԴ տվյալների մշակման ժամանակ: Այս գնահատումները ցույց են տալիս, որ սեղմման մեթոդները ունեն սեղմման տարբեր գործակիցներ և ցուցադրում են յուրահատուկ վարքագիծ տվյալների բաշխված մշակման ընթացքում: ԵԴ տվյալները ներառում են և՛ նկարներ, որոնք նկարահանված են հեռակառավարման տվիչների և արբանյակների կողմից, և՛ տեքստային/թվային տվյալներ տեղային (in-situ) տվիչներից: 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 տվյալներ, բացի այդ նաև կիրառվում են որպես մեծածավալ տվյալների մշակման միջավայրերի արտադրողականության գնահատման հենանիշներ, մեքենայական ուսուցման K-Means ալգորիթմը, որը լայնորեն օգտագործվում է առանձնահատկությունների արդյունահանման, դասակարգման կամ անոմալիաների հայտնաբերման և ԵԴ տվյալների մշակման NDVI ինդեքսը նկարների համար:

Գնահատումները ցույց են տալիս ԵԴ արբանյակային պատկերների համար տվյալների սեղմման մեթոդների կիրառման արդյունավետությունը: Նկար 4-ը ներկայացնում է NDVI ինդեքսի հաշվարկի համար Dask-ի և Spark-ի արտադրողականության համեմատությունը՝ հաշվի առնելով մուտքային տվյալների ծավալները և սեղմման մեթոդները, որոնք աջակցվում են միջավայրերի կողմից:

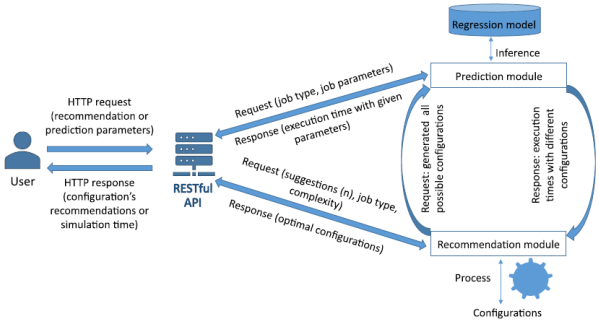


Նկար 4: Dask-ի և Spark-ի համեմատությունը՝ հաշվի առնելով 16, 32, 64 ԳԲ ծավալով մուտքային տվյալները և աջակցվող սեղմման մեթոդները

Գնահատումները ցույց են տալիս, որ Dask-ը և Spark-ը տրամադրում են տվյալների մշակման նմանատիպ կատարման ժամանակ: Dask և Zstandard սեղմման մեթոդների միավորումը տրամադրում է լավագույն արդյունքը, քանի որ սեղմման մեթոդը ապահովում է սեղմման լավագույն գործակիցը բոլոր հնարավոր առանց կորստի սեղմման մեթոդներից: Այն նվազեցնում է օգտագործված պահոցի ծավալը 16%-ով և արագացնում է կատարման ժամանակը 4,72 և 3,99 անգամ համապատասխանաբար Dask-ում և Spark-ում համեմատած Deflate մեթոդի հետ, որը լռելյայնորեն օգտագործվում է որոշ գլոբալ ԵԴ տվյալների պահոցներում:

4.3 Ենթագլուխը ներկայացնում է արտադրողականության օպտիմալացված որոշումների կայացման ծառայությունը, որը բաղկացած է Recommendation և Prediction մոդուլներից: 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

Prediction մոդուլը գնահատում է տվյալների մշակման ժամանակը՝ հիմնվելով տարբեր գծային և բազմանդամ ռեգրեսիոն մոդելների վրա, որոնք ուսուցանվել են սիմուլացիոն տվյալների հավաքածուի 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, β կշիռները սահմանվում են ուսուցման փուլում: X-ը տարբեր հատկանիշներից բաղկացած տվյալների հավաքածուն է, y-ը կատարման ժամանակ է՝ տվյալ X հատկանիշներով, λ-ն կանոնավորացման պարամետրը: 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:

Recommendation մոդուլը հենվելով Prediction մոդուլի տրամադրած արդյունքների վրա, հաշվի առնելով օգտագործողի հարցումը տրամադրում է արդյունավետ առաջարկություն: Մոդուլը դիտարկում է բոլոր հնարավոր կազմաձևերը՝ հաշվի առնելով տվյալների սեղմման աջակցվող մեթոդները տվյալ աշխատանքային հոսքի համար և կազմաձևերը ուղարկում է Prediction մոդուլ՝ տվյալ հնարավոր կազմաձևերի կատարման ժամանակները գնահատելու համար: Վերջապես, Recommendation մոդուլը դասավորում է ստացված արդյունքները ըստ կատարման ժամանակի և վերադարձնում ամենալավ n կազմաձևման պարամետրերը:

4.4 ենթագլուխը ամփոփում է 4-րդ գլխի արդյունքները:

Աշխատանքի հիմնական արդյունքները.

1. Մշակվել է առանց սերվերի ընդլայնվող ԵԴ տվյալների մշակման համալիր համակարգ, որը ճկուն կերպով համատեղում է ԵԴ տվյալների պահեստները ամպային և բարձր արտադրողականությամբ ենթակառուցվածքների հետ [1, 4, 5]:
2. Մշակվել է բազմաֆունկցիոնալ օպտիմալացման մեթոդ՝ հաշվի առնելով մի շարք չափանիշներ, ինչպիսիք են արտադրողականությունը և ծախսերը, ապահովելով միջավայր բաշխված հաշվողական միջավայրերում կլաստերների ընտրության օպտիմալացման համար [3]:
3. Մշակվել է արտադրողականության օպտիմալացված որոշումների կայացման ծառայություն, որն առաջարկում է տվյալների սեղմման արդյունավետ մեթոդներ, հաշվի առնելով ինչպես պահեստային տարածքի խնայողությունը, այնպես էլ տվյալների մշակման արտադրողականության բարելավումը [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.

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.

The research communities and end-users set up local high-performance computing and cloud infrastructures or rely on the resources of global cloud providers to address the increasing needs of EO data processing, which becomes more complex and requires more hardware resources and agile software. There are two approaches to managing EO data: one involves leveraging generic services offered by global resource providers, while the other entails the deployment of specialized platforms. Both approaches have their advantages and limitations. Although offering user-friendly services along with enormous computational and storage capacities, the solutions of the global cloud providers and specialized platforms are closed, therefore, customization and expansion options are constrained, ultimately restricting users to the platforms' generic solutions and diminishing their overall flexibility. Besides that, these solutions often involve charges that users must cover for service access and the vendor lock-in issue is also an actual challenge in this case, since switching to a different solution can be difficult and expensive owing to proprietary formats and tools.

The work aims to deliver a novel EO data processing complex system that overcomes the mentioned limitations and provides flexible and extendable solutions for efficient EO data processing considering crucial key performance indicators.

The purpose and problems of the work

The main purpose of the work is to deliver a scalable and optimized EO data processing complex system combining data repositories and cloud-HPC infrastructures. To achieve this goal, we consider the following problems:

1. Develop a scalable serverless EO data processing system that seamlessly integrates data repositories with cloud-HPC infrastructures, ensuring efficient and flexible processing of EO data.
2. Develop a multi-objective optimization method for distributed computing cluster selection based on various criteria, including performance and cost objectives.
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 EO data processing complex system is developed that flexibly combines EO data repositories with cloud-HPC infrastructures [1, 4, 5].
2. A multi-objective optimization method is developed considering a range of criteria, including performance and cost objectives, providing a framework for optimizing cluster selection in distributed computing environments [3].
3. A performance-aware decision-making service is suggested to recommend efficient compression methods considering both storage space savings and improvements in processing performance [2, 5, 6, 7].

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 и пользователи создают локальные высокопроизводительные вычислительные и облачные инфраструктуры или полагаются на ресурсы глобальных поставщиков облачных услуг для удовлетворения растущих потребностей в обработке данных НЗ, которая становится все более сложной и требует больше аппаратных ресурсов и гибкого программного обеспечения. Существует два подхода к управлению данными НЗ: один предполагает использование общих услуг, предлагаемых глобальными поставщиками ресурсов, а другой влечет за собой развертывание специализированных платформ. Оба подхода имеют свои преимущества и недостатки. Несмотря на то, что решения глобальных облачных провайдеров и специализированных платформ предлагают удобные для пользователя услуги наряду с огромными вычислительными мощностями и мощностями хранения, они закрыты, поэтому возможности настройки и расширения ограничены, что в конечном итоге ограничивает пользователей универсальными решениями платформ и снижает их общую гибкость. Кроме того, эти решения часто предполагают оплату, которую пользователи должны нести за доступ к услугам, и проблема привязки к поставщику также является актуальной проблемой в этом случае, поскольку переход на другое решение может быть трудным и дорогостоящим из-за проприетарных форматов и инструментов.

Целью работы является создание новой комплексной системы обработки данных НЗ, которая преодолевает упомянутые ограничения и обеспечивает гибкие и расширяемые решения для эффективной обработки данных НЗ с учетом важнейших ключевых показателей эффективности.

Цель и рассматриваемые задачи

Основной целью работы является создание масштабируемой и оптимизированной комплексной системы обработки данных НЗ, объединяющей хранилища данных и облачные и высокопроизводительные инфраструктуры. Для достижения этой цели мы рассматриваем следующие проблемы:

1. Разработка масштабируемую бессерверную систему обработки данных НЗ, которая легко интегрирует репозитории данных с облачными и высокопроизводительные инфраструктурами, обеспечивая эффективную и гибкую обработку данных НЗ.

2. Разработка метод многокритериальной оптимизации для выбора кластера распределенных вычислений на основе различных критериев, включая показатели производительности и стоимости.

3. Оценить влияние методов сжатия данных на распределенную обработку больших данных НЗ, обеспечив баланс между экономией хранилища и повышением 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 комплексная система обработки данных НЗ, которая гибко сочетает хранилища данных НЗ с облачными HPC-инфраструктурами [1, 4, 5].
2. Разработан метод многокритериальной оптимизации с учетом ряда критериев, включая показатели производительности и стоимости, обеспечивающий основу для оптимизации выбора кластера в распределенных вычислительных средах [3].
3. Предлагается служба принятия решений с учетом производительности, которая рекомендует эффективные методы сжатия, учитывающие как экономию места для хранения, так и повышение производительности обработки [2, 5, 6, 7].