ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Թոփչյան Արտյոմ Ռուբենի

Տվյալներով պայմանավորված կառավարման նախագծերի իրականացմանը աջակցող ճարտարապետություն ժամանակակից ձեռնարկությունների համար

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

ՍԵՂՄԱԳԻՐ

Երևան –2016

ИНСТИТУТ ПРОБЛЕМ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ НАН РА

Топчян Артем Рубенович

Архитектура для поддержки выполнения проектов, ориентированных на управление по данным, в современных предприятиях

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук по специальности 05.13.04 – «Математическое и программное обеспечение математических машин, комплексов, систем и сетей»

Ереван - 2016

Ատենախոսության թեման հաստատվել է Երևանի պետական համալսարանում

Գիտական ղեկավար՝	ֆիզ. մաթ. գիտ. դոկտոր Ս.Կ. Շուքուրյան
Պաշտոնական ընդդիմախոսներ՝	տեխ. գիտ. դոկտոր Ա. Ք. Պալյան ֆիզ.մաթ.գիտ. թեկնածու Ա.Մ. Վասիլյան
Առաջատար կազմակերպություն՝	Հայաստանի Ազգային Պոլիտեխնիկական Համալսարան

Պաշտպանությունը կայանալու է 2016թ. նոյեմբերի 18-ին, ժ. 16:00-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: Սեղմագիրը առաքված է 2016թ. հոկտեմբերի 18-ին:

Մասնագիտական խորհրդի
գիտական քարտուղար, ֆ.մ.գ.դ. Հ. Գ. Սարուխանյան

Тема диссертации утверждена в Ереванском государственном университете

Научный руководитель:	доктор физ.-мат. наук С. К. Шукурян
Официальные оппоненты:	доктор тех. наук А.Х Палян кандидат физ. мат. наук А.М Василян
Ведущая организация:	Национальный Политехнический Университет Армении

Защита состоится 18-ого ноября 2016г. в 16:00 на заседании специализированного совета 037 «Информатика» Института проблем информатики и автоматизации НАН РА по адресу: 0014, г. Ереван, ул. П. Севака 1.

С диссертацией можно ознакомиться в библиотеке ИПИА НАН РА.

Автореферат разослан 18-ого октября 2016г.

Ученый секретарь специализированного
совета, д.ф.-м.н. А. Г. Саруханян

Արդիականությունը: Տվյալների պահանջարկը և նրանց աճող ծավալները ժամանակակից կազմակերպություններում կարևոր խնդիր են բարձրացրել. ինչպես արդյունավետ օգտագործել արդեն կուտակված և նոր առաջացող տվյալները կազմակերպության գործունեությունը բարելավելու և ավելի արդյունավետ դարձնելու համար: Արձագանքելով այդ տենդեցին ավելանում է այն կազմակերպությունների թիվը, որոնք նպատակ են դնում իրականացնել տվյալներով պայմանավորված կառավարման որոշումների կայացմանը աջակցող նախագծեր[1] [2] և փորձում են աճեցնել տվյալներից քաղվող տեղեկատվության որակը, քանակը, մշակման ու որոշումների կայացման արագությունը, խթանելով կազմակերպության տնտեսական շահը:

Նշված նախագծերը հիմնականում իրականացվում են հետևյալ գործառույթներով՝ բացահայտվում են անհրաժեշտ տվյալների աղբյուրները, հասկացվում է տվյալների բնույթը, ապահովվում է տվյալների հասանելիությունը և վերլուծության իրականացումը: Խնդիր տրիվիալ չէ, եթե տվյալների հնարավոր աղբյուրների քանակը մեծ է, տվյալների կառուցվածքը բաղկացած է շատ բաղադրիչներից և տվյալների մասին տեղեկատվություն պարունակող փաստաթղթերի քանակը շատ արագ աճում է:

Խոշոր կազմակերպություններն ունեն հազարավոր աշխատակիցներ և գործառույթներ, ինչը ծնում է բազմաթիվ համակարգեր, որոնք իրենց հերթին ամսեկան առաջացնում են տասնյակ տեռաբայթ ծավալով տվյալներ: Տվյալները հիմնականում տեղակայված են ռելացիոն հենքերում և ֆայլային համակարգերում, որոնք սփռված են ամբողջ կազմակերպությունով մեկ: Չնայած այն փաստին, որ գոյություն ունեն տեխնիկական համակարգերի օգտագործումը նկարագրող միջոցներ, այնուամենայնիվ չի կարելի պնդել, որ խոշոր կազմակերպություններում կան հստակ նկարագրեր համակարգերում օգտագործվող տվյալների և մետա-տվյալների մասին, որտեղ նկարագրվում է, թե ինչ են պարունակում տվյալների աղբյուրները և ինչպես են օգտագործվում առկա համակարգերում: Այդ մասին է վկայում Google ընկերության մի քանի ամիս առաջ հրապարակված հետազոտությունը[3]: (Հարկ է նշել, որ ներկայացվող աշխատանքում մշակված գործիքակազմը, փաստացի կառուցվել է Google-ի ներկայացված հետազոտություններին զուգահեռ և արդյունքների համեմատությունը դժվար է

[1] I. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. *Data wrangling: The challenging journey from the wild to the lake.* In CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2015.

[2] Rahman, Nayem, and Fahad Aldhaban. *"Assessing the effectiveness of big data initiatives."* 2015 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE. 2015.

[3] Halevy, Alon, et al. *"Goods: Organizing Google's Datasets."* Proceedings of the 2016 International Conference on Management of Data. ACM, 2016.

իրականացնել, քանի որ հրապարակման մեջ բացակայում են տեխնիկական մանրամասները)։

Պրակտիկայում տվյալների համակարգեր մշակելիս առաջնային դեր չի հատկացվում տվյալների տեռաբայթների, հազարավոր կտոր փաստաթղթերի, աղյուսակների, տիպերի և այլ մետատվյալների օգտագործումը նկարագրող փաստաթղթերի ստեղծմանը։ Դա է պատճառը, որ խնդիր է առաջանում, երբ կազմակերպությունում անհրաժեշտ է լինում կարճ ժամանակում կառուցել ծրագրային փոփոխություն, որը արագ կարձագանքի հաճախորդի ձեռնարկած գործողությանը, հատկապես եթե պահանջվում է հատուկ լուծում՝ օրինակ կարճատև հապաղում ապահովող, որն արդեն կաշխատի վայրկյաններ նախորդ լուծման ժամերի փոխարեն։

Իհարկե գոյություն ունեն համակարգեր, որոնք փորձում են արձագանքել նշված խնդիրներին, օրինակ հայտնի են տվյալների հասանելիությունն ապահովող Data Lake և Data Warehouse[4] համակարգերը և գիտելիքի կառավարման համար Enterprise Data Management[5] համակարգերը։ Բայց նորից կարելի է նկատել, որ բացակայում է միասնական ճարտարապետության մոտեցումը, որի նպատակը կլիներ միասնական լուծում տալ բարձրացված խնդրին, հատկապես այնպիսի կազմակերպությունների տեսակետից, որոնք արդեն ունեն տեղակայված ու տարիներ շարունակ գործածվող կազմակերպչական կառուցվածք ու տարատեսակ համակարգեր և որոնք հետևում են CRISP-DM[6] մոդելին, որը սահմանում է տվյալների պեղման և օգտագործման մոտեցումները։

Ատենախոսության նպատակը։ Ատենախոսության նպատակն է կառուցել ճարտարապետություն և մի շարք գործիքներ, որոնք միասին հասցեագրված կանդրադառնան տվյալներով կառավարվող նախագծերի ներդրման խնդրին։ Այս գործիքները և ճարտարապետությունը պետք է բավարարեն կազմակերպության հատուկ պահանջներին (օրինակ պետք է բավարարեն CRISP-DM մոդելին), լինեն ընդլայնվող, անխափան, սահմանափակ ռեսուրսներ օգտագործող և ապահովեն կարճատև հապաղման պայմանը։ Ինչպես նաև պետք է ընդգրկվեն հետևյալ գործառույթները՝

- **Տվյալների փնտրման** և տվյալների մասին տեղեկատվության փոխկապակցվածությունների բացահայտումը ապահովող ենթահամակարգ, որը հիմնվում է մետատվյալների, ֆունկցիոնալ նկարագրերի և փաստաթղթերի վրա։

- **Տվյալների փորձարարական միջավայր։** Բոլոր կազմակերպական տվյալների հավաքագրման և մշակման ենթահամակարգ, որը ապահովում է տվյալներով

[4] Patil, Preeti S and Rao, Srikantha and Patil, Suryakant B. *Optimization of data warehousing system: Simplification in reporting and analysis.* IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET); 9:33-37, 2011.

[5] Berson, Alex, and Larry Dubov. *Master data management and customer data integration for a global enterprise.* McGraw-Hill, Inc., 2007.

[6] Shearer C. *The CRISP-DM model: the new blueprint for data mining.* J Data Warehousing; 5:13—22. 2000

կառավարվող նախագծերի մշակողների հետ համատեղ և համագործակցային կառուցում, հեշտացնում է նման նախագծերի նկարագրումը և մեկնարկումը։

Հետազոտական մեթոդները, որոնք օգտագործվել են ատենախոսությունում ներառում են բնական լեզվի մշակման մեթոդները, որոնք հիմնված են հավանականային թեմատիկ մոդելների վրա, բանալի արտահայտությունների առանձնացման ալգորիթմները։ մեծ ծավալի տվյալների համեմատության և հարցման ալգորիթմները։ փնտրման շարժիչների կառուցման մոտեցումները։ համակարգային և բաշխված ծրագրավորումը։ բաշխված համակարգերի ճարտարապետության նախագծումը։ տվյալների հոսքերի մշակումը, տվյալների հավաքագրման, անխափան և կարճատև հապաղումով պրոցեսների էվոլյուցիոն մասշտաբավորումը։

Գիտական նորույթը: Ատենախոսությունը պարունակում է հետևյալ նոր արդյունքները.

1. Առաջարկվում է մեծ քանակի փոխկապակցված՝ կառուցվածքով և առանց կառուցվածքի, տվյալների աղբյուրների միջավայրում էությունների և կապերի բացահայտմանը աջակցող փնտրման համակարգի ճարտարապետություն։

2. Առաջարկվում է տվյալներով պայմանավորված նախագծերի CRISP-DM մոդելի չափանիշներին համապատասխան փորձարարական միջավայրի իրականացման ճարտարապետություն։

3. Առաջարկվում է ճարտարապետության միասնական իրականացում, որն ապահովում է տվյալների աճի պայմաններում մասշտաբավորում, կարճատև հապաղում, տվյալների զրոյական կորուստ և համակարգի նվազ սպասարկում միջավայրի պարամետրերի փոփոխության պայմաններում։

Կիրառական նշանակությունը և ներդրումները: Նկարագրված լուծումը մշակված և գործարկված է MAN GmbH (http://www.truck.man.eu/) ներքին ցանցում և ամպային տվյալների կենտրոններում։ Այն ակտիվորեն օգտագործվում է և դրական գնահատական ստացել շուրջ 20 տարբեր թիմերի կողմից, որոնք աշխատում են տվյալներով կառավարվող նախագծերի մշակման վրա MAN ընկերությունում։ Լուծման հարթակի ինտեգրումից հետո, բազմաթիվ մասնակիցների կողմից նշվել է իրենց սովորական աշխատանքի արտադրողականության նկատելի աճ։ Մասնավորապես, պահանջվող տվյալների ու փաստաթղթերի հասանելիության ժամանակը բարելավվել է (աղյուսակ 1)։

Լուծման հարթակը գտնվում է ակտիվ մշակման փուլում և թիմերի հետ համագործակցելով ավելացվել են նոր հնարավորություններ։ MAN ընկերությունում պլանավորվում է նաև մշակված հարթակի նոր գործարկումներ բազմաթիվ հավելյալ միջավայրերում։

	Սովորաբար առանց հատուկ գործիքակազմի	ODP համակարգի կայուն տարբերակի միջավայրում
Տվյալների համախմբին դիմելու ժամանակը	2 ամիս	2 շաբաթ
Տվյալների համախմբի բովանդակության ծանոթացում	3 շաբաթ	1 շաբաթ
Նախագծի մշակման միջավայրի հասանելիությունը	1 ամիս	1 շաբաթ
Նախագծի մասնակից վերլուծողներ	3	20
Ակտիվ նախագծեր	2	10

Աղյուսակ 1.

Լուծման հարթակը ներդրված է Data Reply GmbH (http://www.reply.com/) կազմակերպությունում և ակտիվ օգտագործվում է, ներկայումս քննարկվում է օգտագործման հնարավորություն Գերմանիայի այլ խոշոր կազմակերպություններում:

Պաշտպանության են ներկայացվում հետևյալ դրույթները.

1 Տվյալների փնտրման համակարգը, որը՝

- ❖ Ստեղծում է կազմակերպչական հասկացությունների տաքսոնոմիա:
- ❖ Կառուցվում է փաստաթղթերում օգտագործված տվյալների աղբյուրների հենք, հիմնվելով կազմակերպչական մետատվյալների վրա, օրինակ համակարգերի անվանումները, աղյուսակների սխեմաները, սյուները և այլն:
- ❖ Ճանաչում և բացահայտում է փաստաթղթերի, նախագծերի և տվյալների աղբյուրների փոխկապակցվածությունները:
- ❖ Կառուցում է շարունակաբար նորացվող ինդեքսներ, այնպես որ հասկացությունների ու կառուցվածքների փոփոխությունները անհապաղ արտացոլվում և կիրառվում են:

2 Տվյալների փորձարարական միջավայրը, որն՝

- ❖ Ապահովում է կազմակերպչական հիմնարար տվյալների ամբողջականությունը:
- ❖ Ապահովում է առկա կազմակերպչական տվյալների կապերի մոդելավորման սահմանված միջոց:

- ❖ Հարմարվում է կազմակերպության տվյալների համակարգի փոփոխություններին՝ չազդելով ընթացքի մեջ գտնվող այլ նախագծերի վրա:
- ❖ Ապահովում է լիարժեք մեկուսացված, փոխարինելի և ինքնանկարագրվող տվյալներին դիմելու անվտանգ և համագործակցային միջավայր:

3. Հիմնվելով աշխատանքում ներկայացված մոտեցումների և արդյունքների վրա մշակվել է միասնական Գործառնական Տվյալների Հարթակ (այսուհետև ODP): Ճարտարապետությունը մշակված է արդյունաբերական ստանդարտներին համապատասխան և պատրաստ է գործարկման խոշոր կազմակերպություններում: ODP հարթակի իրականացումը ապահովում է՝
   - ❖ Հոսքային մշակման համակարգերի միջոցով Կարճատև հապաղման պայմանի կատարումը, տեղեկատվությանը դիմելիս և նորոգելիս:
   - ❖ Հարմարեցումը ներառվող տվյալներին, սպառողներին և պահանջում է սպասարկման նվազ կարիք:
   - ❖ Անխափան աշխատանք և հասանելիության բարձր աստիճան:

Ապրոբացիա և հրապարակումներ: Ատենախոսության հիմնական արդյունքները ներկայացվել և քննարկվել են ԵՊՀ SS Հետազոտական և Կրթական կենտրոնի ընդհանուր սեմինարներին, ՌԴ Գիտությունների Ակադեմիայի Համակարգային ծրագրավորման ինստիտուտում, Data Reply GmbH կազմակերպության և Google Germany GmbH համատեղ տարբեր աշխատաժողովներում և սեմինարներում (10.07.2015), ինչպես նաև MAN GmbH, Volkswagen GmbH և LIDL GmbH (1.8.2016) կազմակերպություններում: Աշխատանքը զեկուցվել է Data Reply GmbH կազմակերպության International Xchange Conference համաժողովներում 2015 և 2016թթ.: Ստացված արդյունքները և մշակված մոտեցումները ներկայումս կիրառվում են MAN GmbH կազմակերպությունում:

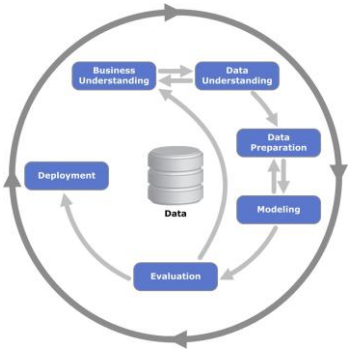
Աշխատանքի արդյունքներն արտացոլված են թվով վեց հոդվածներում, որոնց ցանկը բերված է այս սեղմագրի վերջում:

Աշխատանքի կառուցվածքը և ծավալը: Ատենախոսությունը բաղկացած է ներածությունից, 4 գլուխներից, եզրակացությունից և օգտագործված շուրջ 65 հղումների ցանկից: Հիմնական տեքստը պարունակում է 105 էջ:

Ատենախոսության բովանդակությունը

Ներածությունը անդրադառնում է ատենախոսության թեմայի արդիականությանը և պրակտիկ կիրառելիությանը արդյունաբերության համատեքստում: Քննարկվում է ոլորտում առկա միջազգային լավագույն փորձը նման խնդիրների հաղթահարման համար և սահմանվում են ատենախոսության հիմնական նպատակը, հիմնական արդյունքները և մոտեցման նորույթը, ինչպես նաև առաջարկվող լուծման կիրառական արժեքը:

Առաջին գլուխը մանրամասնորեն նկարագրում է, թե ինչ են իրենցից ներկայացնում տվյալներով կառավարվող նախագծերը, ինչպես են ղեկավարվում արդյունաբերությունում և ներդրման ժամանակ հանդիպող խնդիրները բիզնես ստորաբաժանումների, SS բաժնի և տվյալների հետազոտողի տեսակետից:



Նկար1: Տվյալների պեղման համար Cross Industry Standard Process (CRISP)

Բաժին 1.1-ում մանրամասն նկարագրվում է, թե ինչ է տվյալներով կառավարվող նախագիծը և նրա կարևորությունը այն խոշոր կազմակերպության համատեքստում, որոնք նպատակ ունեն իրենց համակարգերը օգտագործել հնարավորինս արդյունավետ: Նկարագրվում են գոյություն ունեցող մոտեցումներ նման նախագծերի կառավարման և ուսումնասիրման համար: Հիմնական շեշտը դրվում է տվյալների պեղման Cross Industry Standard Process (CRISP) մոտեցման վրա, որն իրեն լավ է ցուցաբերել արդյունաբերության ոլորտում՝ նման նախագծերի համակարգման և կառավարման համար: Նկարագրում է այդ մոտեցման կիրառելիությունը տվյալներով կառավարվող նախագծեր մշակող հայտնի խոշոր կազմակերպություններում հանդիպած հիմնական դեպքերի համար:

Բաժին 1.2-ում նկարագրվում են կազմակերպական և տեխնիկական խնդիրները, որոնք ծագում են տվյալներով կառավարվող նախագծերի իրականացման ժամանակ: Հատուկ ուշադրություն է դարձվել աշխատաժամանակի ծախսի վրա, որին կարող է բերել մշակման ժամանակ CRISP ստանդարտի փուլերին հետևելը: Արդյունքները ընդհանրացված են Նկար

1-ում: Դիտարկվում են խնդիրները բիզնեսի ստորաբաժանումների, SS բաժնի և տվյալների հետազոտողների թիմի տեսանկյուններից: Վերջիններս հանդիսանում են տվյալներով կառավարվող նախագծերի հիմնական շահակիցները: Խնդիրները խիստ կապված են ներգրավված ստորաբաժանումների ցանկի հետ: Հատկապես՝

- SS բաժինը առնչվում է տվյալների կառավարման և տեխնիկական հարցերին, ինչպիսիք են տվյալների մեծ ծավալի համաձայնեցված կառավարումը և այդպիսի ծավալների մշակման իրականացումը:
- Բիզնեսի ստորաբաժանումներ, որոնք ունեն թափանցիկության պակաս հատկապես նրանում, թե որ տվյալները և քանի տվյալների աղբյուր կարող են միավորվել, որպեսզի իդենդիֆիկացվի պոտենցիալ բիզնես դեպքը:
- Տվյալների հետազոտողների թիմը պայքարում է տվյալներ և տվյալների մասին տեղեկատվություն հայթայթելու համար, ինչը կարող է օգտակար լինել բիզնեսի կողմից բացահայտված խնդիրը լուծելու համար:

Բաժին 1.3-ում ուրվագծվել են շահակիցների կողմից ընդունվող և օգտագործվող մոտեցումներ, որոնք անդրադառնում են նշված հարցերին: Հատկապես նշվում է՝

- Ինչպես են Apache Hadoop կլաստերը և Data Lake մոդելները [7] օգտագործվում տվյալների կազմակերպման և մշակման համար:
- Ինչպես է Enterprise Data Management գործիքը [8] օգտագործվում տվյալների փոխկապակցվածությունը առավել խորը հասկանալու համար:
- Ինչպես են Documentation Management և Enterprise Search համակարգերը օգտագործվում տվյալների, նախագծերի և դրանց բովանդակության մասին տեղեկություն հայթայթելու համար:

Քանի որ այդ ծրագրային լուծումները լավ են աշխատում միայն որոշակի դեպքերում մենք դուրս է բերել, թե ինչում է կայանում անհրաժեշտությունը ունենալ տրամաբանորեն հիմնավորված միասնական լուծում, որը կարող է հասցեագրվել բոլոր երեք դեպքերին միաժամանակ: Նշվում է, թե ինչպես այդ փաստը կդարձնի տվյալներով կառավարվող նախագծի իրականացման ամբողջական գործընթացը ավելի արդյունավետ:

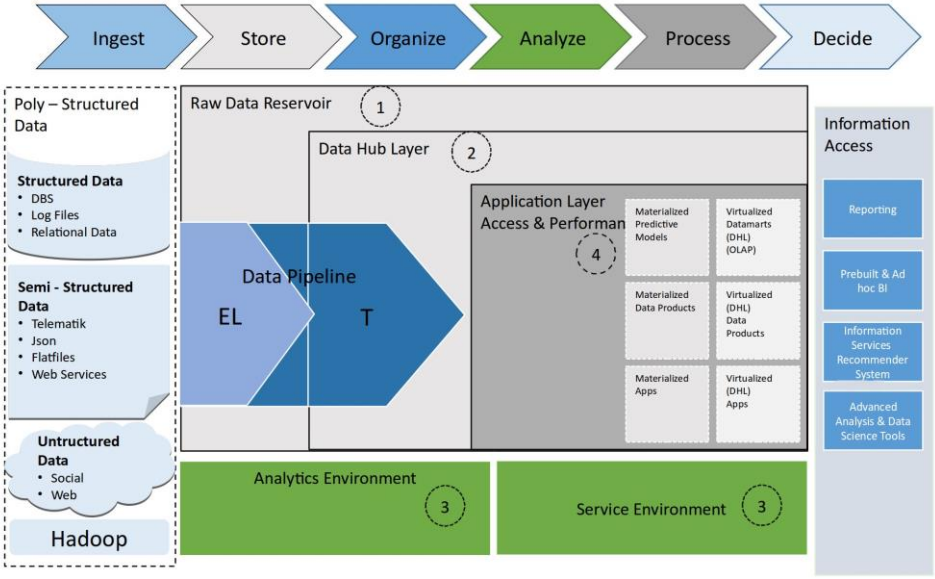
Երկրորդ գլխում անդրադառնվում է առաջարկվող լուծմանը տվյալների պահոցի կառավարման և մշակման համար: Նաև նկարագրվում է, թե մշակման ինչպիսի միջավայրեր է առաջարկվում բազմասպասարկման (multi-tenancy), ծառայության տրամադրման

[7] I. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. *Data wrangling: The challenging journey from the wild to the lake. In CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, 2015

[8] Berson, Alex, and Larry Dubov. *Master data management and customer data integration for a global enterprise.* McGraw-Hill, Inc., 2007.

(provisioning) և մասշտաբավորման կառավարման հարցերը տվյալների հետազոտողների և ծրագրերի մշակողների կողմից լուծելու համար:

Վերջինս արվում է առաջին հերթին վերլուծելով նախորդ գլխում նկարագրված ստանդարտ լուծումներում ծագած խնդիրները և ուրվագծում է այն պահանջների ցանկը, որոնք հասցեագրվում են նշված խնդիրների լուծմանը: Այս ցանկը կազմվել է հիմնվելով մշակումների և բազմաթիվ կազմակերպություններում առկա փորձի վրա: Այնուհետև անդրադառնվում է առաջարկվող լուծմանը, որը փորձում է արձագանքել նշված պահանջներին: Լուծումը հատկապես կենտրոնանում է տվյալների հավաքագրման, պահեստավորման և մշակման համար նախատեսված, հստակ սահմանված, մասշտաբավորվող, աննշան ուշացումով մոդելների ստեղծման վրա և նշվում է, թե ինչպես են պարզեցվում տվյալներին հասանելիությունը, թիմային համագործակցությունը, մշակումը, թեստավորումը և գործարկումը:



Նկար 2: Տվյալների անոթ/Data reservoir

Բաժին 2.1-ում նկարագրվում է առաջարկվող լուծման տրամաբանական և կառուցվածքային մոդելը, որը հիմնված է Data Vault մոդելավորման մոտեցման վրա [9]: Նշվում են այս մոդելի առավելությունները և առաջարկվող մոդելի տարբերությունները

[9] Dan Linstedt. *Super Charge your Data Warehouse*. ISBN 978-0-9866757-1-3. 2010.

արդյունաբերությունում օգտագործվող այլ մոդելներից, ինչպիսիք են՝ Data Lakes, Data Warehouses: Առաջարկվող մոդելի տրամաբանական կառուցվածքը ընդհանուր գծերով տրվում է Նկար 2-ում:

Բաժին 2.2-ում նկարագրվում են տեխնիկական խնդիրներ, որոնց հանդիպել ենք նկարագրված մոդելի իրականացման ժամանակ, հատկապես մասշտաբավորման, կարճատև հապաղումով և կառուցվածքի պահանջներին բավարարելու համար: Հատուկ նկարագրվում է թե ինչպես է մոդելավորել տվյալների ընկալման և հիմնական պահոցի մակարդակները՝ դիտելով տարբեր համակարգերում պահված տվյալները որպես կառուցվածքով տվյալների միասնական հոսք՝ ի հակադրություն փաթեթային մշակման մոտեցման մեջ ընդունված մոտեցմանը: Արդյունքում նկարագրվում է Lambda [10] ճարտարապետությունը և նրա վարիացիաները, որոնց վրա է հիմնված առաջարկվող մոտեցումը: Տրվում է տվյալների հոսքի և իրական ժամանակում մշակման վրա հիմնված նմանակ մոդելի կիրառման առավելությունը փաթեթային մշակման նկատմամբ: Այնուհետև ընդգծվում է, թե ինչպես է առաջարկվող լուծումը ընդլայնում վերը ներկայացվածը՝ լիովին հիմնվելով հուսալի ու կարճատև հապաղումով տվյալների զտման և պահպանման օգտագործվող հոսքերով մոտեցման վրա:

Բաժին 2.3-ում տրվում է բաժին 2.2-ում նկարագրված մոդելի տեխնիկական ճարտարապետությունը և իրականացումը: Սահմանվում և նկարագրում է այդպիսի համակարգ կառուցելու համար անհրաժեշտ կոմպոնենտները: Հատուկ ուշադրություն է դարձվում այնպիսի կոմպոնենտների վրա, որոնք անհրաժեշտ են հոսքերի վրա հիմնված մասշտաբավորվող, ամուր և հուսալի համակարգերի կառուցման համար: Նշվում է, թե ինչպես են առաջարկվող լուծման մեջ ներառվում այլ համակարգեր ինչպիսիք են Apache Kafka-ն [11], որը հուսալիորեն ներկայացնում է տվյալների հոսքերը, Apache Mesos-ը [12], որը ապահովում է կիրառվող տվյալների մշակման հոսքագծերի անխափանելիությունը, մասշտաբավորումը և առանձնացված կոմպոնենտային կառավարումը, Kafka-Connect-ը [13], որը օգնում է տվյալների զտման և տվյալների բեռնման հուսալի հանգույցների մասշտաբավորվող հիմքի կառուցմանը: Մանրամասնորեն դիտարկում է, թե ինչպես է ապահովվում տվյալների կլանման գործընթացի հուսալիությունը, ամեն տվյալի միայն մեկ անգամ մատակարարման սկզբունքի պահպանումը և ավտոմատ մասշտաբավորումը, որը հիմնվում է մշակման համար օգտագործվող հաշվողական ռեսուրսների և մշակվող տվյալների քանակի էվոլյուցիոն մեթոդներով կատարվող անալիզի արդյունքների վրա[14]:

[10] Marz, Nathan, and James Warren. *Big Data: Principles and best practices of scalable realtime data systems.* Manning Publications Co. 2015

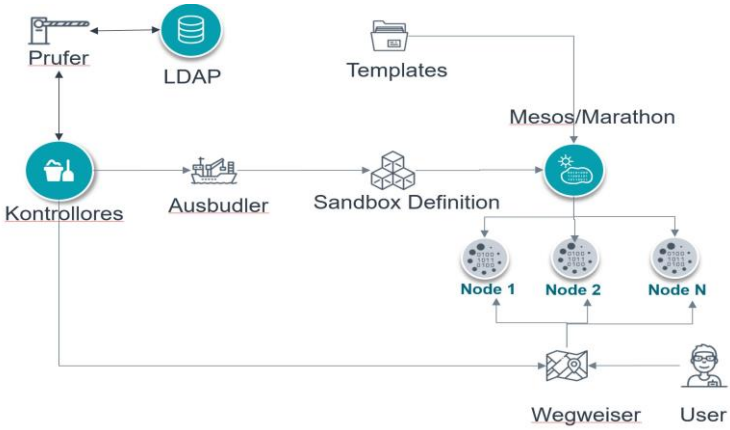
[11] Dunning, Ted, and Ellen Friedman. *Streaming Architecture: New Designs Using Apache Kafka and Mapr Streams*. O'Reilly Media 2016.

[12] *Apache Mesos* http://mesos.apache.org 2015.

[13] *Kafka-Connect.* http://docs.confluent.io/2.0.0/connect. 2015.

[14] Newell, Andrew, et al. *"Optimizing distributed actor systems for dynamic interactive services."* Proceedings of the Eleventh European Conference on Computer Systems. ACM. 2016

Բաժին 2.4-ում նկարագրվում է առաջարկվող լուծման տեխնիկական ճարտարապետությունը և իրականացումը: Սահմանում և նկարագրում են համակարգը կառուցելու կոմպոնենտները: Մասնավորապես ուրվագծվում է, թե ինչպես է օգտագործվել միկրոծառայությունների[15] մոդելը՝ մշակման միջավայրի կյանքի ցիկլը ավտոմատացնող համակարգ ստեղծելու համար: Այս միջավայրը կարող է շատ պարզեցնել տվյալների հետազոտողի և ծրագրերի մշակողի աշխատանքը: Լուծման կոմպոնենտների ուրվագիծը տրված է Նկար 3-ում:



Նկար 3: Փորձնական միջավայրի ուրվագիծ

Գլուխ 3-ում նկարագրվում է, թե ինչպես է առաջարկվում հասնել տեղեկատվության փնտրման և գիտելիքի փոփոխության՝ վերը նկարագրված խնդրի լուծմանը: Ինչպես նկարագրվում է նախորդ գլխում առաջին հերթին ավելի մանրամասն ուրվագծվում է տվյալներով կառավարվող նախագծերի կառուցման տարածված իրականացումները և առաջարկվում է պահանջների ցանկ, որին պետք է բավարարեն այդպիսի համակարգերը՝ առավել արդյունավետ լուծում տրամադրելու համար: Ցանկը ստեղծված է մեծ թվով կազմակերպությունների հետ համագործակցելով և հավաքագրելով նրանց արձագանքը առկա մարտահրավերներին: Հիմնվելով առաջադրված պահանջների վրա սահմանվում է նորարարական մոտեցում տեքստերի և մետատվյալների վերլուծության համար: Ուրվագծվում է թե ինչ մեթոդներ և մոդելներ են օգտագործվել մեծ քանակի փաստաթղթերից պահանջվող տեղեկատվությունը զտելու և դուրս բերելու համար: Նկարագրվում են պահանջվող տեղեկատվությունը զտելու համար կիրառվող ալգորիթմները, որոնք կառուցվել են պահպանելով որդեգրված մասշտաբավորման և կարճատև հապաղման ոճը:

[15]* Newman, Sam. *Building Microservices* . O'Reilly Media, Inc, 2015.

Բաժին 3.1-ում ավելի մանրամասն նկարագրվում են կազմակերպությունների կողմից առաջադրված խնդիրները՝ տվյալների և բիզնես փոխըմբռնման համատեքստում: Նշվում են խնդիրները, որոնք առաջանում են տվյալների կառավարման դասական մոտեցումների կիրառման ժամանակ: Ներկայացվում է պահանջների ցանկ որոնք պետք է հասցեագրվեն այդ խնդիրներին: Պահանջների ցանկը հավաքագրված է հիմնվելով կուտակված փորձի և առաջարկվող լուծման նախագիծը օգտագործող մի շարք մեծ կազմակերպությունների հետ համագործակցության վրա:

Բաժին 3.2-ում տրամադրված են մանրամասներ, թե ինչպես է սահմանվում տեղեկատվության փնտրման և վերլուծելու խնդիրը՝ որպես փաստաթղթերի ցանկի ձևափոխելու և հարստացնելու փուլեր, որոնցով անցնում է փաստաթուղթը նախքան փնտրման համակարգում ինդեքսավորվելը և տեքստային հարցումների համար հասանելի դառնալը: Մանրամասն նկարագրվում են փուլերը և դրանց արտապատկերումը սահմանված պահանջներին:

Բաժին 3.3-ում ուրվագծվում է, թե ինչ մոտեցումներ և մոդելներ են օգտագործվում վերլուծության փուլերում, որպեսզի հնարավոր դառնա պահանջվող տեղեկատվության զտումը: Գլուխը կենտրոնանում է հետևյալ թեմաների վրա.

- Ինչպես է կիրառվում թեմատիկ մոդելավորումը , հատկապես Latent Dirichlet Allocation-ը [16] և նրա տարբերակները, փաստաթղթերից լավ սահմանված համատեքստային ներկայացումը զտելու համար:
- Տեքստի ամփոփման ալգորիթմներ, օրինակ Textrank[17] ալգորիթմը, որը կիրառվում է փաստաթղթի տեքստից սեղմագիր արտահանելու համար:
- Օգտագործելով տեղայնորեն զգայուն[18] (locally sensitive) հեշավորման ալգորիթմը և օգտվելով կազմակերպության տվյալների աղբյուրի սահմանումներից, տվյալների հենքերի աղյուսակներից և այլ մետատվյալներից դուրս բերել, թե փաստաթղթերի բազմության որ միավորներում կան հղումներ դեպի հայտնի տվյալներ, տվյալների աղբյուրներ և նախագծեր:

Բաժին 3.4-ում նկարագրվում է առաջարկվող տեխնիկական ճարտարապետությունը, նրա իրականացումն և ալգորիթմները: Տրվում են մանրամասներ թե ինչպես է առաջարկվող լուծումը իրականացված է գլուխ 2-ում նկարագրված հոսքերի մասշտաբավորվող մշակման մոտեցմամբ: Նկարագրում է, թե ինչպես է ավտոմասշտաբավորումը, խափանումներից վերականգնումը և ռեսուրսների կառավարման վերը ներկայացված գաղափարները միասին օգտագործվում ենթահամակարգի իրականացման համար: Նկարագրվում են տեխնիկական

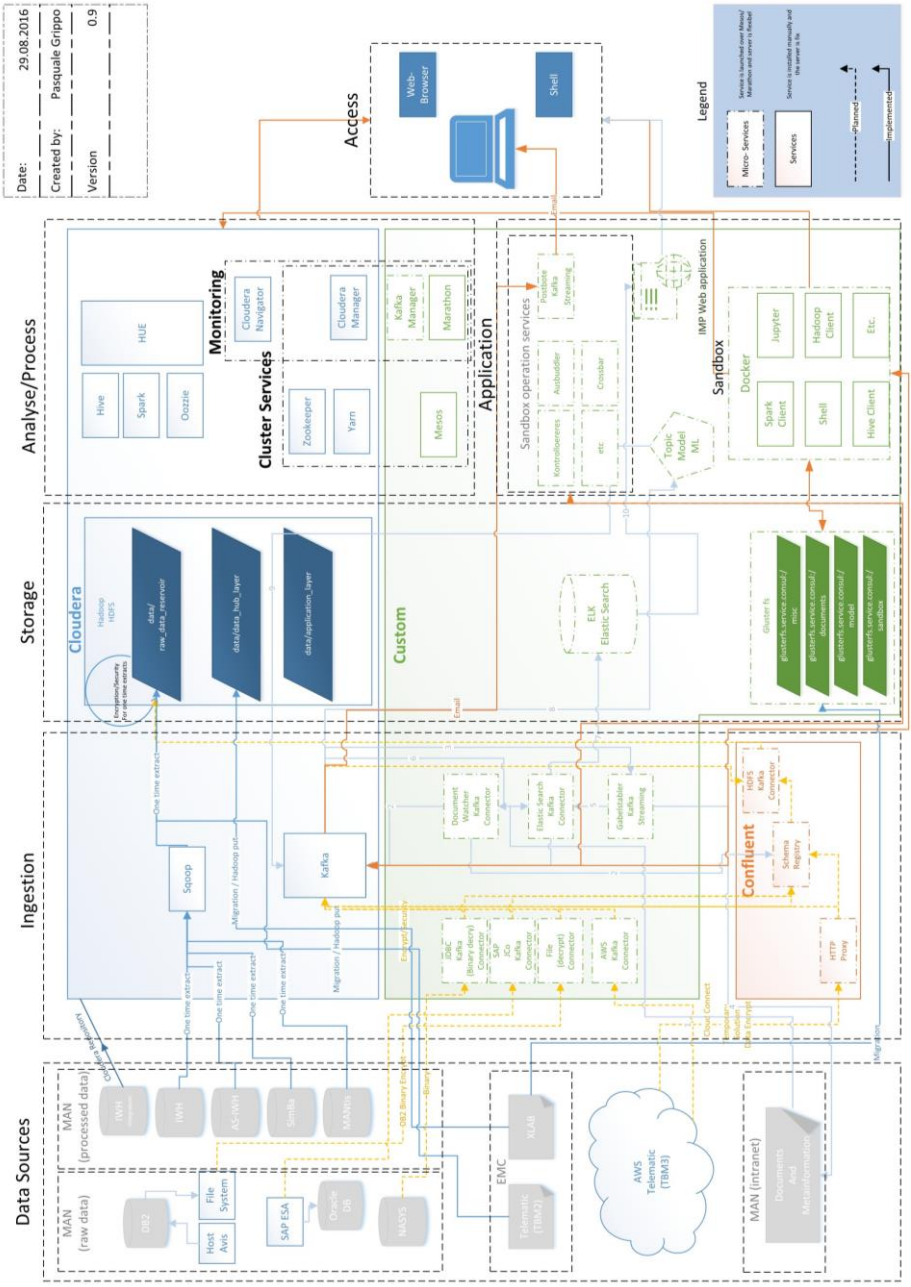
[16] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. *"Latent dirichlet allocation."*Journal of machine Learning research 3.Jan: 993-1022. 2003

[17] Mihalcea, Rada, and Paul Tarau. *"TextRank: Bringing order into texts."* Association for Computational Linguistics. 2004.

[18] E. Cohen et al. *"Finding interesting associations without support pruning."* IEEE Transactions on Knowledge and Data Engineering, vol. 13, no. 1, pp. 64-78. 2001.

հարցեր, որոնք հանդիպել են թվարկված մոդելները կիրառելիս, այն դեպքում երբ մեծ կազմակերպություններում առկա են տարբեր տիպի համակարգեր, մեծ թվով փաստաթղթեր, մեծ քանակի տվյալներ, մետատվյալներ, օրինակ աղյուսակներ, աղյուսակների սյուների անուններ և այլն: Հատուկ ուշադրություն է դարձվել կարճատև հապաղման և մասշտաբավորման պահանջների վրա:

Գլուխ 4-ում նկարագրում է թե ինչպես են առաջադրվող բոլոր լուծումները ձևավորում միևնույն ճարտարապետական պրիմիտիվների վրա հիմնված միասնական հարթակ: Լուծումը սահմանվում է, որպես միասնական Գործառնական Տվյալների Հարթակ, որը նպատակ ունի տրամադրել անխափան և մասշտաբավորվող մոտեցում՝ մեծ կազմակերպություններում տվյալներով կառավարվող նախագծերի առավել արդյունավետ իրականացման համար: Ուրվագծված ճարտարապետությունը ներկայացված է նկար 4-ում:



Պատկեր. 4: Գործառնական տվյալների պլատֆորմ

Ատենախոսության հիմնական արդյունքները

1. Առաջարկվում է տվյալներով կառավարվող նախագծերի իրականացման համար կազմակերպության տեղեկատվական աղբյուրների հենքի կոնցեպտուալ մոդել, որը բավարարում է նախագծերի պահանջներին, ապահովում է տվյալների սկզբնաղբյուրների ու տեղեկատվությունը զտելու, պահպանելու և մշակելու միջոցների հասանելիությունը՝ հաշվի առնելով կազմակերպության առկա կազմակերպական համակարգերը և սահմանափակումները:
2. Կազմակերպության մետատվյալների և փաստաթղթերում առկա գիտելիքների հենքի հետախուզման համար կառուցվել է տեղեկատվության փնտրման միջավայր: Մասնավորապես փաստաթղթերի ամփոփման, խմբավորման և կոնտեքստային նշագրման համար, ինչպես նաև կազմակերպության տվյալների, իրականացվող նախագծերի և դրանց փոխկապակցությունների բացահայտելու համար օգտագործվել են տեքստի անալիզի մոտեցումները:
3. Առաջարկվում է «Փորձարարական» միջավայր, որ բազմադիսցիպլինար թիմերին հասանելի է դարձնում կազմակերպության տվյալներն ու հաշվողական ռեսուրսները և օգտագործվում է նոր տվյալների մոդելներ հետազոտելու և գնահատելու համար: Այսպիսով թիմերին տրամադրում է անվտանգ միջավայր և ապահովվում է հիմնական կազմակերպչական համակարգերի համար ամբողջականության պայմանը ռեսուրսների նրանց հասանելիության մեկուսացման միջոցով:
4. Պլատֆորմը իրականացվել է հոսքային մշակման, ռեսուրսների կառավարման և միկրոծառայությունների գաղափարների վրա հենվելով, ինչը թույլ է տալիս արդյունավետ օգտագործել սահմանափակ տեխնիկական ռեսուրսները, ամեն անգամ տեղայնացնելով և հարմարեցնելով պլատֆորմը առկա կազմակերպական համակարգերին և պրոցեսներին: Հիմնվելով դրա վրա պլատֆորմը ունի հնարավորություն թափանցիկ մասշտաբավորվելու ինչպես կազմակերպության տվյալների ծավալին, այնպես էլ օգտագործողների քանակին համապատասխան:
5. Պլատֆորմի ներդրումը կարճ ժամանակում առաջին մոտարկմամբ արդեն ցույց է տվել է ժամանակի տնտեսման տենդենց տվյալներով պայմանավորված շուրջ 10 նախագծեր մշակող թիմերի 20 աշխատակիցների գործառույթներում 3-ից 4 անգամ:

Դիսերտացիայի թեմայով հրապարակումներ

[1] *Artyom Topchyan.* Enabling Data Driven Projects for a Modern Organization. Proceedings of the Institute for System Programming, Volume 28 (Issue 3), 2016.

[2] *Artyom Topchyan.* Information Retrieval and Analysis for a Modern Organization. Proceedings of the Institute for System Programming, Volume 28 (Issue 4), 2016.

[3] *Artyom Topchyan.* Scalable Sandbox Environments for a Modern Organization. Proceedings of the Institute for System Programming, Volume 28 (Issue 4), 2016.

[4] *Artyom Topchyan.* Scalable Documentation Search Engine. Proceedings of Crisis Management and Technologies, Volume 10, 2016.

[5] *Artyom Topchyan.* Platform for Data Driven Projects. Proceedings of Crisis Management and Technologies, Volume 10, 2016.

[6] *Artyom Topchyan, Tigran Topchyan.* Muscle-based skeletal bipedal locomotion using neural evolution. In Proceedings of IEEE, Computer Science and Information Technologies (CSIT), pages 1–6, 2013.

Resume

Artyom Topchyan

ARCHITECTURE ENABLING DATA DRIVEN PROJECTS FOR A MODERN ORGANIZATIONS

With the growing volume and demand for data a major concern for an Organization is how to use this data more effectively to generate value for the organization. To address this, more and more Organizations are aiming at implementing Data-Driven projects [19] [20], which are increasing the quality, speed, and/or quantity of information gain for innovating a new methodology or the economic benefit to an organization.

These projects are about finding data, understanding data and accessing it. This has become not a simple task with the amount of data and documentation being created at organizations growing rapidly. Large organizations have thousands of employees that create dozens of systems, which produce 10's of TB's of data every month. All of this data is stored in database- and file-systems scattered throughout the organization. While there is a defined way to manage the descriptions of such technical systems themselves, the same thing is not true for all the data and meta-data which describe what the data actually contains, which is confirmed by the research carried out by Google 21. The outlined tool, which was developed in parallel with the presented research and the first publication, which does not contain technical details was published a few months ago.

TB's of data and thousands of documentation pieces and meta-data definitions for tables, types and etc do not have a defined way of finding and using them in projects. This has become even a larger problem, with the growing requirement for more low-latency use-cases, where Organizations want to very quickly react to something a customer does. Low-Latency also means reacting in seconds as opposed to hours or days, which is commonly the case. While there are systems, that approach these issues, such as Data Lakes and Data Warehouses[22] for data access and Enterprise Data Management[23] Systems for knowledge management, there is no single architectural approach that aims to efficiently solve these problems together from the specific view point of an Organization

[19] I. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. *Data wrangling: The challenging journey from the wild to the lake.* In CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2015.

[20] Rahman, Nayem, and Fahad Aldhaban. *"Assessing the effectiveness of big data initiatives."* 2015 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE. 2015.

[21] Halevy, Alon, et al. *"Goods: Organizing Google's Datasets."* Proceedings of the 2016 International Conference on Management of Data. ACM, 2016.

[22] Patil, Preeti S and Rao, Srikantha and Patil, Suryakant B. *Optimization of data warehousing system: Simplification in reporting and analysis.* IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET); 9:33-37, 2011.

[23] Berson, Alex, and Larry Dubov. *Master data management and customer data integration for a global enterprise.* McGraw-Hill, Inc., 2007.

with years of organizational structure and system already in place and follows the CRISP-DM model[24], which defines the methodlogy for search and using data.

The aim of the dissertation is to create an Architecture and number of tools which address the problem of creating Data-Driven projects at an Organization. These Architecture and tools have to address the specific requirements of organizations: they have to support the CRISP-DM model, be scalable, fault-tolerant, functional under limited resource contains and support low-latency processing. The following components are supported:

- A flexible way for discovering data and interconnections of the data, based on meta-data, functional descriptions and documentation.
- System for collecting and processing all of the organizational data, that allows for collaborative and shared creation of Data-Driven projects as well simplify testing and deployment of said projects.

Main results of the work

1. Developed a conceptual model for creating Data-Driven projects on the basis of an Organizations data, which fulfill the project requirements, provides access to the raw data and to sources of data ingestion, storage and processing capabilities, while taking into account the existing organizational systems and restrictions.
2. Proposed an Information Retrieval environment for exploring the organizations metadata and knowledge basis contained in documentation. Text Analysis approaches for summarizing, context labelling and clustering documents are used to discover information about the organizational data and projects as well as uncovering relationships between them.
3. Developed "Sandbox" environments, which provide multi-disciplinary teams access to the organizations data and computational resources, which are used to research new data models and evaluate them. These environments provide security to the teams and ensure the integrity of the source organizational systems by means of full resource and access isolation.
4. The platform is implemented based on the concepts of stream processing, resource management and micro-services, which allows to effectively utilize limited hardware resources, customize and configure the platform to fit the existing organizational systems and processes. Based on this the platform can transparently scale to the volume of the organizations data as well as a large number of users.
5. Based on a first approximation, using the platform led to the shortening of processing time about 3 to 4 time during the first 5 months immediately after system deployment. Data is based on the first integrated data sources and 20 developers working on this data in 10 projects.

[24] Shearer C. *The CRISP-DM model: the new blueprint for data mining.* J Data Warehousing ; 5:13—22. 2000

Заключение

Артем Топчян

Архитектура для поддержки выполнения проектов, ориентированных на управление по данным, в современных предприятиях

С ростом объема и новых запросов на данные одной из важных задач для организаций становится эффективное использование накопившихся и поступающих данных с целью создания дополнительных преимуществ для самих организаций. Реагируя на это обстоятельство, все больше организаций нацеливаются на внедрение проектов, ориентированных на управление по данным [25] [26]. Одновременно делаются попытки, направленные на увеличение качества и скорости выделения информации из поступающих данных, так и увеличения количества выделяемой информации. Одновременно стимулируется сам процесс быстрого принятия решений в пользу экономической выгоды организаций. Такие проекты реализуют процессы обнаружения, понимания и обеспечения доступа к необходимым данным, что становится не простой задачей с учетом растущего количества самих данных и различной документации, создаваемой вокруг этих данных. Крупные организации, имеющие тысячи сотрудников, создают десятки систем, производящих десятки тербайтов данных каждый месяц. Обычно все эти данные хранятся в базах и файловых системах, разбросанных по всей организации. В то время как организации серьезно поддерживают пользовательские описания управления техническиими системами, они не так основательно занимаются описанием самих данных и мета-данных, описывающих, какой контекст несут в себе базовые данные. Наличие такой тенденции подчеркивается свежими исследованиями, проведенными и опубликованными компанией Google. (Представлемый нами инструмент, был разработан параллельно с упомянутыми исследованиями Google, однако и первая их публикация, которая не содержит технические деталей, вышла всего несколько месяцев назад.)

Терабайты данных, тысячи томов документации, определения метаданных для таблиц, типов данных и т.д. не имеют единого метода поиска с целью их использования в проектах, нацеленных на анализ данных. Это становится еще большей проблемой в связи с увеличением случаев, когда организации в очень короткий срок необходимо подготавить новое программное решение, например, реагирующее на действия клиентов в реальном времени. Задача еще более усугубляется растущей потребностью в программах,

[25] I. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. *Data wrangling: The challenging journey from the wild to the lake.* In CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 2015.

[26] Rahman, Nayem, and Fahad Aldhaban. *"Assessing the effectiveness of big data initiatives."* 2015 Portland International Conference on Management of Engineering and Technology (PICMET). IEEE. 2015.

обеспечивающих более низкие задержки обработки. Существуют системы, которые касаются этих вопросов, например, озера данных и хранилища данных[27], обеспечивающие управление доступом к данным, и системы управления данными предприятий [28], нацеленные на управление знаниями. Однако не существует единого архитектурного подхода, который призван эффективно решать упомянутые проблемы в совокупности, с учетом специфики отдельных организаций, у которых в наличии устоявшаяся организационная структура, действующие системы обработки данных и которые придерживается специальных правил, например, внедряют модель CRISP-DM[29], поддерживающую определенную методологию поиска и использования данных.

Целью диссертационной работы является построение архитектуры и инструментария, нацеленных на задачи реализации проектов, ориентированных на управление по данным. Инструментарий и архитектура должны удовлетворять особым тербованиям организаций, например, соответствие CRISP-DM модели, будут масштабируемыми, отказоустойчивыми, используют ограниченные ресурсы и обеспечивают требования малой задержки обработки данных. Одновременно должна быть обеспечена следующая функциональность:

- поиск данных и обнаруживание информации о взаимосвязях данных на основе метаданных, функциональных описаний и документов:
- опытной среды – “песочницы”, обеспечивающей возможность сбора и обработки организационных данных, а также обеспечивать согласованное взаимодействие групп разработчиков в процессе совместного построения программных решений для проектов, ориентированных на управление по данным.

Основные результаты работы

1. Предлагается концептуальная модель базы информационных источников организации, используемых для создания проектов, ориентированных на управление по данным. Она отвечает требованиям проекта, обеспечивает доступ как к исходным данным, так и самим способам выделения, хранения и обработки данных, учитывает существующие организационных системы и присущие им ограничения.
2. Предлагается информационно-поисковая среда для выявления метаданных и базы знаний, содержащихся в документациях. Для выделения информации об организационных данных и проектах, а также обнаружения связей между ними применяется подход анализ текстов, позволяющий обобщать, контекстно-маркировать и кластеризовать документы.

---

[27] Patil, Preeti S and Rao, Srikantha and Patil, Suryakant B. *Optimization of data warehousing system: Simplification in reporting and analysis.* IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET); 9:33-37, 2011.

[28] Berson, Alex, and Larry Dubov. *Master data management and customer data integration for a global enterprise.* McGraw-Hill, Inc., 2007.

[29] Shearer C. *The CRISP-DM model: the new blueprint for data mining.* J Data Warehousing; 5:13—22. 2000

3. Предлагается среда построения «песочниц» анализа данных, которые обеспечивают доступ мультидисциплинарных команд к данным организации и ее вычислительным ресурсам в процессе построения новых моделей данных и их оценки. Такие среды обеспечивают командам безопасность и целостность исходных организационных систем посредством полной изоляции используемых ресурсов и ограничения доступа к ним.
4. Платформа реализована на основе концепций обработки потоков данных, управления ресурсами и микро-услуг, что позволяет эффективно использовать ограниченные аппаратные ресурсы, настраивать и конфигурировать платформу, под существующие организационные системы и процессы. Исходя из этого платформа может прозрачно масштабироваться по объему данных организации, а также при увеличении числа пользователей.
5. Внедрение платформы, за первые 5 месяцев экслпуатации, показало тенденцию экономии затраты времени, в первом приближении, от 3-х до 4-х раз для 10-и проектов с вовлечением 20-и сотрудников.