

**АНАЛИТИЧКА ПЛАТФОРМА ЗА ПОДРШКУ ПРИСТУПУ РАЗВОЈА СОФТВЕРА
ВОЂЕНОМ ПОДАЦИМА****ANALYTICS PLATFORM FOR THE SUPPORT OF DATA-DRIVEN APPROACH OF
SOFTWARE DEVELOPMENT**

Тамара Лазаревић, Факултет техничких наука, Нови Сад

**Област – ПРИМЕЊЕНЕ РАЧУНАРСКЕ НАУКЕ
И ИНФОРМАТИКА**

Кратак садржај – У овом раду представљена је имплементација, интеграција са примером апликације у оквиру модела софтвера као услуге (енг. SaaS), као и евалуација аналитичке платформе за подршку развоја софтвера вођеном знањем. Платформа је развијена у циљу повећања изгледа за успешан пласман производа и повећања конкурентске предности компанија на тржишту. Пројектована је модуларно, при чему су дати конкретни алати за имплементацију наведених модула. Улаз у платформу представљају подаци који потичу из интеракције корисника са апликацијом, док је излаз скуп дашборда који омогућавају информисано доношење одлука о даљем развоју апликације. Резултати евалуације потврдили су да решење испуњава захтеве за реалну употребу.

Кључне речи: аналитике, платформа, data-driven, развој софтвера, рударење знања

Abstract – This paper presents the implementation, integration with an example application within the Software as a Service (SaaS) model, and the evaluation of an analytical platform for supporting data-driven software development. The platform aims to boost the chances of successful product launch and enhance companies' competitive advantage in the market. It is modular, with specific tools provided for implementing the stated modules. The platform's input consists of data derived from user interactions with the application, while the output is a set of dashboards that enable informed decision-making about the further development of the application. The evaluation results confirmed that the solution meets the requirements for real-world use.

Keywords: analytics, platform, data-driven, software development, data mining

1. УВОД

Адекватна примена компанијског знања, знања о производу и корисницима кључна је за конкуритивну предност компанија [1], [2].

НАПОМЕНА:

Овај рад проистекао је из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

Истраживања истичу важност повезивања корисничког знања са инжењерским процесима за успешан пласман производа на тржиште [2].

Успех компанија у којима су одлуке вођене подацима (енг. data-driven) зависи од ефикасне примене знања на производе који доносе вредност корисницима [2], па кључан проблем представља одрживо и исплативо прикупљање и интеграција тог знања у развојне процесе.

У раду је дат предлог реализације платформе за подршку дата-дривен приступу развоја софтвера.

Примена платформе илустрована је на случају пример софтвер као услуга (енг. SaaS, Software as a service) апликације за организацију посла која прати канбан методологију, пружајући комплетан графички увид у пројекте.

Подаци који фигуришу у платформи представљају синтетичке податке који би настали потенцијалним корисничким коришћењем апликације.

Представљена је модуларна архитектура платформе, описана интеграција са постојећим компонентама апликационог система, као и крајњи излаз из платформе и његови бенефити. Дати су и конкретни алати који се могу користити за имплементацију модула платформе, и њихове предности.

Евалуација платформе обављена је тестирањем перформанси неколико модула под редовним и вишим интензитетом коришћења, употребом алата за тестирање система под оптерећењем. Добијени резултати указују на то да платформа може успешно да одговори на оба типа оптерећења и неометано обавља своју функционалност, и да моделовање података знатно утиче на перформансе платформе.

2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ**2.1. Примена знања о корисницима у процесу развоја софтвера**

Ране студије релевантне за проблематику истичу значај управљања знањем у компанијама [1]. Каснија истраживања се фокусирају и на важност разумевања потреба и понашања корисника за постизање конкурентске предности и финансијског успеха компанија [2].

У раду [3] истичу се предности коришћења техника машинског учења и рударења знања (енг. *data mining*) за компаније: за идентификацију првих корисника, фокусирања комуникације на препознати сегмент купаца, предвиђања будуће потражње и откривања оптималних побољшања за производ или услугу.

Радови из последњих 5 година потврђују значај знања о корисницима, и позивају на интеграцију знања о корисницима у процес развоја производа [4]. Изучавањем је и утицај дигитализације и растуће улоге података у процесу дизајна и развоја производа, с акцентом на академска питања која произилазе из тога. Ова питања обухватају откривање могућих концептуалних модела, процедура и платформи за прикупљање података [5], што је обрађено у раду.

2.2. Технике рударења знања за екстракцију знања о корисницима

У литератури која се бави употребом техника рударења знања за екстракцију знања о корисницима, коришћене технике рударења знања укључују асоцијативна правила, кластеровање, стабла одлучивања (енг. *decision trees*), мајновање знања из текста (енг. *text mining*), концоинт анализу (енг. *conjoint analysis*) и друге, и употребљавају се у различитим фазама развоја производа [4].

Извори података о корисницима у литератури су најчешће резултати интернет анкетања [2], [4]. Резултујући скупови података могу представљати ограничен узорак са малом варијансом, што може утицати на биас резултата. Овај рад насупрот, предлаже методологију прикупљања знања директно од корисника, насталог њиховом употребом и интеракцијом са апликативним производом.

2.3. Системи за управљање подацима

Рад [6] дели архитектуре аналитичких платформи у три групе. Прва група обухвата архитектуре у којима складиште података има централну улогу, где су мане мањак подршке за неструктуриране податке и високе трошкове платформи за велике количине података. Модернији приступ су двослојне архитектуре које укључују слој језера података (енг. *data lake*). Подаци се чувају у слојевима са ниском ценом складиштења, а само подскуп података се трансформише и даље транспортује у складишта података за аналитичке примене. Према раду, користе се у готово свим Форчн 500 компанијама. Недостаци приступа су сложеност, слабљење поузданости у податке, слаба свежина података, слаба подршка напредних аналитичких метода као што је машинско учење, висока цена решења. Рад предлаже лејкхаус (енг. *lakehouse*) архитектуру као решење, засновану на отвореним форматима за директан приступ подацима, подршком за машинско учење и науку о подацима, и најсавременијим перформансама. Међутим, ова технологија доноси изазове као што су недостатак стручног кадра и зависност од једног пружаоца услуга.

3. ТЕОРИЈСКИ ПОЈМОВИ И ДЕФИНИЦИЈЕ

3.1. Дата-дривен менаџмент

Дата-дривен менаџмент се ослања на доношење одлука заснованих на анализи и интерпретацији података. Компаније са овом стратегијом ослањају се на податке у циљу персонализовања производа и изградње комплетног односа са перспективама и клијентима [7].

3.2. ОЛАП

ОЛАП (енг. *Online Analytical Processing*) системи су класа система за управљање базама података. Оптимизовани су за аналитичке, агрегационе упите са мањим бројем колона над великом количином историјских података, за генерисање аналитичких извештаја и дашборда. Подаци се ређе ажурирају, и не представљају податке у реалном времену. Насупрот томе, ОЛТП системи управљају континуалним током трансакција, константно мењајући податке. Налазе се у многим системима са којима корисници директно интерагују, попут веб апликација [8].

3.3. ЕТЛ и ЕЛТ

Кључни појмови за разумевање ЕТЛ и ЕЛТ процеса су:

- **Extract** (Екстракција) - добављање података из апликације и других сервиса
- **Transform** (Трансформација) - чишћење, форматирање, агрегирање, обогаћивање, комбиновање и организовање података
- **Load** (Учитавање) - Складиштење података на одредишту

ЕТЛ се често користи као појам вишег нивоа апстракције, под који спадају и ЕТЛ и ЕЛТ - учитавање, складиштење и трансформација података. Кључна разлика је што се код ЕТЛ-а трансформација података врши ван складишта података. Код ЕЛТ-а, подаци се трансформишу након учитавања. ЕЛТ постаје стандард како су складишта постала моћнија и јефтинија [9].

3.4. Принципи инкременталног и потпуног освежавања

Приступ потпуног освежавања (енг. *full refresh*) у сваком извршавању трансформација или екстракције поновно процесуира комплетни садржај изворне табеле (или табела), што резултује новом резултујућом табелом. Инкрементални (енг. *incremental*) приступ проналази редове табеле који су креирани или измењени од последње итерације, и обрађује само те редове. Инкрементални приступ, иако сложенији, је најкориснији за обраду великих табела, када је потпуно процесуирање неефикасно [10].

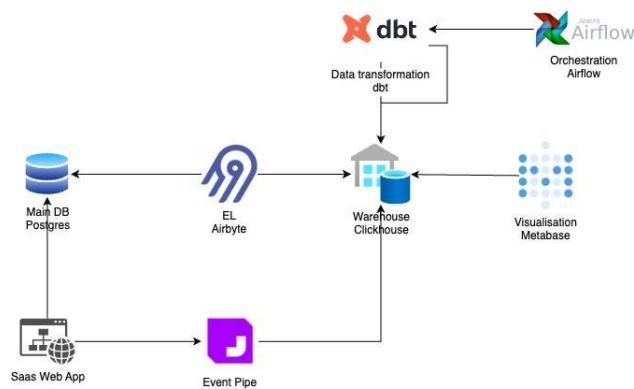
3.5. Аналитике засноване на догађајима

Аналитике засноване на догађајима (енг. *event-based*) служе за праћење и анализирање интеракција између корисника и производа. Догађај је свака акција или понашање које се дешава унутар дигиталне тачке контакта, као што су мобилна апликација, имејл или веб страница. Корисници представљају потенцијалне клијенте који ступају у интеракцију са компанијом у дигиталном окружењу. Када корисници комуницирају са производом, бележи се њихово понашање и кориснички подаци, као што су локација, уређај и

језик. Ова комбинација пружа детаљније информације о томе како и зашто се производ користи.

4. МЕТОДОЛОГИЈА

Дијаграм архитектуре аналитичке платформе представљен је на слици 1.



Слика 1. Дијаграм архитектуре система

Архитектура платформе подељена је на модуле: ЕЛ и канал за догађаје за прикупљање података, модуле за складиштење, трансформацију, визуелизацију података и модул за оркестрацију. Платформа је организована у модуларну хијерархију *docker compose* конфигурационих фајлова, што омогућава лако управљање и замену модула. *Docker compose* је алат за дефинисање и извршавање вишеслојних апликација.

4.1. ЕЛ модул

ЕЛ модул као улогу има аутоматизовану екстракцију података са извора и њихово допремање на одредиште. У конкретном случају платформе извор података представља база података софтвер као услуга апликативног система. Централизованом одредиште за податке представља складиште података.

За имплементацију ЕЛ модула коришћен је *Airbyte* [11].

4.2. Канал за догађаје модул

Модул канал за догађаје прати догађаје настале током корисничког коришћења апликације, функционишући као посредник између извора догађаја и одредишта. Улазни догађаји се прикупљају са клијентске или сервер стране, групишу, трансформишу и чувају у систему за складиштење података. За имплементацију модула коришћен је *Jitsu* [12]. Догађаји могу бити подразумевани или специјално дефинисани. У примеру апликације дефинисан је и имплементиран скуп догађаја од интереса за даљу обраду.

4.3. Модул за складиштење података

Модул за складиштење података омогућава централизацију података из различитих извора на једном месту и њихову интеграцију.

За имплементацију изабран је *ClickHouse*. *ClickHouse* представља добар избор за потребе платформе због своје специфичне архитектуре која пружа оптималне перформансе у ОЛАП сценаријима [13].

4.4. Модул за трансформацију података

Модул за трансформацију података врши интеграцију података из различитих извора у трансформисане податке, погодне за коришћење у аналитичке сврхе. Улаз у модул су подаци различитих извора, а излаз су трансформисани подаци којима крајњи корисници платформе имају приступ.

Модул је имплементиран као независна компонента у архитектури система. Кључна компонента и алат овог модула је *dbt*. *dbt* је алат који заузима место стандарда у индустрији за трансформацију података [14].

4.5. Модул за оркестрацију

Модул за оркестрацију развијен је за програмирање, заказивање и праћење радних токова (енг. *workflows*). Обавља задатак редовне интеграције података, уз могућност праћења и отклањања грешака.

Као алат за подршку процесу оркестрације изабран је *Apache Airflow* [15], који је скалабилан и омогућава динамичко и прилагодљиво писање радних токова.

4.6. Модул за визуелизацију

Модул за визуелизацију представљен је платформом *Metabase* [16] са скупом дашборда и корисничким налозима са одговарајућим правима приступа. Два глава дашборда пружају увид у коришћење апликације: Клијенти и зарада који садржи основне метрике које омогућавају увид у здравље базе претплатника, као и Поглед у резултате кластеровања који представља визуелизоване резултате потенцијалне сегментације корисника апликације методом кластеровања. Овај модул је тачка платформе са којом интерагују крајњи корисници платформе - развојни тим, који може укључивати: *product owner*-а, маркетинг тим, тим за продају, вођу тима и друге.

4.7. Интеграција са платформом

Апликације засноване на моделу софтвера као услуге морају се интегрисати са аналитичким платформама. Без обзира на механизам прикупљања података, потребно је прилагодити постојеће апликације за коришћење њихових података у аналитици, што захтева и сарадњу развојних тимова. За илустрацију процеса интеграције са платформом реализована је пример апликација и представљене неопходне измене које укључују додавање нових зависности апликације, измену њеног изворног кода и прилагођавање базе података ЕЛ алатима.

5. ЕКСПЕРИМЕНТИ

Сprovedена је евалуација неколико компоненти платформе: модула за складиштење података са утицајем адекватне имплементације модула за трансформацију на перформансе, као и евалуација модула који представља канал за догађаје. Тестиране су перформансе у просечним и условима високог интензитета коришћења.

У оквиру оба наведена експеримента коришћен је алат *Apache Jmeter* [17]. Резултати тестирања модула за складиштење под вишим оптерећењем од 10 упита од стране 10 паралелних корисника, приказани су на слици 2, изражени у милсекундама и агрегирани за све упите.

Label	# Samples	Average	Min	Max	Std. Dev.	Error %	Throughput	Received ...	Sent KB/sec	Av. Bytes
BI Query ...	56235	169	3	1739	169.33	0.00%	464.7/sec	24.28	0.00	53.5
TOTAL	56235	169	3	1739	169.33	0.00%	464.7/sec	24.28	0.00	53.5

Слика 2. Резултати тестирања дешборд упита над оптимално моделованом табелом

Евалуација модула који представља канал за догађаје показује да успешно обрађује захтеве 200 паралелних корисника и допрема догађаје до складишта. Резултати су приказани на слици 3.

Label	# Samples	Average	Min	Max	Std. Dev.	Error %	Throughput	Received ...	Sent KB/sec	Av. Bytes
HTTP Request	200	5	1	38	5.03	0.00%	20.1/sec			
TOTAL	200	5	1	38	5.03	0.00%	20.1/sec			

Слика 3. Резултати за евалуациону групу од 200 нити

6. ЗАКЉУЧАК

У раду је представљена модулarna платформа која пружа подршку дата-дживен приступу у развоју софтверских решења, имплементирана у циљу унапређења вредности софтверских решења за крајње кориснике и повећања шансе позитивног исхода приликом пласмана на тржиште.

Евалуација платформе спроведена је над неколико модула платформе, кроз експерименте који су симулирали различите типове оптерећења. Тестирања су потврдила поузданост и ефикасност платформе, доказујући да решење испуњава захтеве за реалну употребу.

Побољшање у односу на постојећу литературу представља коришћење реалних података о корисницима и њиховим навикама. Предложена побољшања укључују подршку напреднијих аналитичких техника, проширење скупа дешборда и метрика које се посматрају. Са инфраструктурног становишта, у зависности од потреба компаније, разликују се и оптималне архитектуре решења, па би се модулarna архитектура решења могла у складу са њима прилагодити или проширити.

7. ЛИТЕРАТУРА

- [1] C. Soo, T. Devinney, D. Midgley, и A. Deering, „Knowledge Management: Philosophy, Processes, and Pitfalls“, *Calif. Manage. Rev.*, том 44, изд. 4, стр. 129–150, Јули 2002, doi: 10.2307/41166146.
- [2] C.-T. Su, Y.-H. Chen, и D. Y. Sha, „Linking innovative product development with customer knowledge: a data-mining approach“, *Technovation*, том 26, изд. 7, стр. 784–795, Јули 2006, doi: 10.1016/j.technovation.2005.05.005.
- [3] L. Bstieler и остали, „Emerging Research Themes in Innovation and New Product Development: Insights from the 2017 PDMA-UNH Doctoral Consortium“, *J. Prod. Innov. Manag.*, том 35, изд. 3, стр. 300–307, Мај 2018, doi: 10.1111/jpim.12447.
- [4] Y. Zhan, K. H. Tan, и B. Huo, „Bridging customer knowledge to innovative product development: a data mining approach“, *Int. J. Prod. Res.*, том 57, изд. 20, стр. 6335–6350, Окт. 2019, doi: 10.1080/00207543.2019.1566662.
- [5] M. Cantamessa, F. Montagna, S. Altavilla, и A. Casagrande-Seretti, „Data-driven design: the new challenges of digitalization on product design and development“, *Des. Sci.*, том 6, стр. e27, 2020, doi:

10.1017/dsj.2020.25.

- [6] M. Armbrust, A. Ghodsi, R. Xin, и M. Zaharia, „Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics“, 2021.
- [7] SYDLE, „Data-Driven: What It Is and Why It’s Important“, Blog SYDLE. Приступљено: 19. Септембар 2024. [На Интернету]. Available at: <https://www.sydle.com/blog/data-driven-what-it-is-and-why-it-s-important-606c8a4e4b136c41e0e2c334>
- [8] „What Is Online Transactional Processing (OLTP)? | IBM“. Приступљено: 19. Септембар 2024. [На Интернету]. Available at: <https://www.ibm.com/topics/oltp>
- [9] „ETLs, ELTs, and Reverse ETLs“. Приступљено: 19. Септембар 2024. [На Интернету]. Available at: <https://www.metabase.com/learn/grow-your-data-skills/analytics/etl-landscape>
- [10] „Incremental models in-depth | dbt Developer Hub“. Приступљено: 19. Септембар 2024. [На Интернету]. Available at: <https://docs.getdbt.com/best-practices/materializations/4-incremental-models>
- [11] „Airbyte | Open-Source Data Movement for LLMs | AI Platform“. Приступљено: 25. Септембар 2024. [На Интернету]. Available at: <https://airbyte.com/>
- [12] „Jitsu“. Приступљено: 25. Септембар 2024. [На Интернету]. Available at: <https://jitsu.com/>
- [13] ClickHouse, „Fast Open-Source OLAP DBMS“, ClickHouse. Приступљено: 13. Септембар 2024. [На Интернету]. Available at: <https://clickhouse.com>
- [14] „dbt Labs Builds Momentum as the Industry Standard for Data Transformation“, dbt Labs. Приступљено: 15. Септембар 2024. [На Интернету]. Available at: <https://www.getdbt.com/blog/dbt-labs-builds-momentum-as-the-industry-standard-for-data-transformation>
- [15] „Home“, Apache Airflow. Приступљено: 25. Септембар 2024. [На Интернету]. Available at: <https://airflow.apache.org/>
- [16] „Metabase | Business Intelligence, Dashboards, and Data Visualization“. Приступљено: 25. Септембар 2024. [На Интернету]. Available at: <https://www.metabase.com>
- [17] „Apache JMeter - Apache JMeter™“. Приступљено: 17. Септембар 2024. [На Интернету]. Available at: <https://jmeter.apache.org/>

Кратка биографија:



Тамара Лазаревић рођена је у Шапцу 1997. год. Смер Рачунарство и аутоматика на Факултету техничких наука у Новом Саду уписује 2016. године. Основне студије завршава 2020. године. Исте године потом, на истом факултету уписује мастер студије. контакт: tamaralazarevic@uns.ac.rs