

Анализа тржишта електронске трговине

Архитектуре система великих скупова података

1. Домен

Електронска трговина омогућава компанијама и појединцима да врше куповину и продају производа или услуга путем интернета, без физичке присутности на локацији на којој се ти производи или услуге нуде. Како је интересовање за ову грану привреде расло, тако се проширивао и њен домен. Уопштено говорећи, у домен електронске трговине могу се сврстати све активности које корисници предузимају током њиховог искуства на одређеној интернет платформи. Ради поједностављања домена пројекта, разматраће се само акције везане за прегледање производа, додавање или брисање производа из дигиталне корпе, као и вршење трансакција.

2. Мотивација

Због своје практичности, доступности и глобалног досега, електронска трговина доживљава експоненцијалан раст у протеклим годинама. Експанзија ове индустрије довела је до значајног повећања конкуренције на тржишту. У оваквом окружењу, компаније и појединци приморани су да пронађу иновативне начине за издвајање и опстанак на тржишту. Временом, у пракси се показало да разумевање понашања корисника током процеса куповине директно утиче на успешност електронског пословања. Управо зато, циљ ове анализе је боље разумевање овог понашања ради побољшања корисничког искуства, а самим тим и успешности интернет продаје.

3. Пакетна обрада података

Централна идеја пакетне обраде историјских података је стицање увида у понашање корисника током интеракције са системом, конкретно приликом извршавања трансакција. Анализом овог понашања, стиче се увид у потенцијалне препреке које могу ометати корисника у његовој намери да успешно обави куповину. У наредним секцијама налази се опис скупа података који ће бити коришћен за пакетну обраду, као и неколико упита на које пакетна обрада треба ефикасно да одговори.

3.1 Опис скупа података

Скуп података, приказан на слици 1, преузет је са сајта *Kaggle.com*. Изабрани скуп садржи податке који описују понашање корисника приликом интеракције са системом. Подаци су прикупљени током седам месеци, конкретно у периоду од октобра 2019. до априла 2020. године. Подаци потичу из компаније која се бави интернет продајом производа из различитих категорија. Сваки ред у датотеци представља догађај. Сваки догађај је у вези са корисником и производом. Поменути веза је типа *више према више*.

event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2019-11-01 00:00:00 UTC	view	1003461	2.05301355563188E+018	electronics.smartphone	xiaomi	489.07	520088904	4d3b30da-a5e4-49df-b1a8-ba5943f1dd33
2019-11-01 00:00:00 UTC	view	5000088	2.05301356610087E+018	appliances.sewing_machine	janome	293.65	530496790	8e5f4f83-366c-4f70-860e-ca7417414283
2019-11-01 00:00:01 UTC	view	17302664	2.0530135538535E+018		creed	28.31	561587266	755422e7-9040-477b-9bd2-6a6e8fd97387
2019-11-01 00:00:01 UTC	view	3601530	2.05301356381078E+018	appliances.kitchen.washer	lg	712.87	518085591	3bfb58cd-7892-48cc-8020-2f17e6de6e7f
2019-11-01 00:00:01 UTC	view	1004775	2.05301355563188E+018	electronics.smartphone	xiaomi	183.27	558856683	313628f1-68b8-460d-84f6-cec7a8796ef2
2019-11-01 00:00:01 UTC	view	1306894	2.05301355892022E+018	computers.notebook	hp	360.09	520772685	816a59f3-f5ae-4ccd-9b23-82aa8c23d33c
2019-11-01 00:00:01 UTC	view	1306421	2.05301355892022E+018	computers.notebook	hp	514.56	514028527	df8184cc-3694-4549-8c8c-6b5171877376
2019-11-01 00:00:02 UTC	view	15900065	2.05301355819041E+018		rondell	30.86	518574284	5e6ef132-4d7c-4730-8c7f-85aa4082588f
2019-11-01 00:00:02 UTC	view	12708937	2.0530135535599E+018		micellin	72.72	532364121	0a899268-31eb-46de-898d-09b2da950b24
2019-11-01 00:00:02 UTC	view	1004258	2.05301355563188E+018	electronics.smartphone	apple	732.07	532647354	d2d3d2c6-631d-489e-9fb5-06f340b85be0
2019-11-01 00:00:03 UTC	view	17200570	2.05301355979263E+018	furniture.living_room.sofa		437.33	518780843	aa806835-b14c-45af-9538-cd4d1849be3e
2019-11-01 00:00:03 UTC	view	2701517	2.05301356391144E+018	appliances.kitchen.refrigerators		155.11	518427361	c89b0d96-247f-4044-9c91-bb5f38c6af9b
2019-11-01 00:00:03 UTC	view	16700260	2.05301355990168E+018	furniture.kitchen.chair		31.64	566255262	173d7b72-1db7-4638-8207-be8148bf3c9f
2019-11-01 00:00:04 UTC	view	34600011	2.06098132058191E+018			20.54	512416379	4dfe2c67-e537-4dc2-ae69-0b65263db091
2019-11-01 00:00:05 UTC	view	4600658	2.05301356394499E+018	appliances.kitchen.dishwasher	samsung	411.83	526595547	aab33a9a-29c3-4d50-84c1-8a2bc9256104
2019-11-01 00:00:06 UTC	view	24900193	2.05301356218339E+018			1.09	512651494	f603c815-f51a-46fe-9404-cb586e35edaf

Слика 1. Део скупа података намењеног за пакетну обраду.

3.2 Упити

1. На основу броја прегледа производа, исписати информације о најатрактивнијој категорији.
2. Приказати информације о пет главних потрошача, као и број производа који су купили.
3. На основу броја продатих производа, одредити пет марки које најмање доприносе укупној заради.
4. Приказати информације о највише и најмање прегледаном производу.
5. Приказати информације о категорији чији су производи највише пута уклоњени из корпе.
6. Упоредити успешност две категорије производа. Првој категорији припадају производи који имају цену у опсегу од 200 до 400 динара, док другој припадају они чија се цена креће од 400 динара па навише. Успешност категорија дефинисана је бројем продатих производа.

7. Утврдити да ли, и колико постојање марке утиче на продају. Другим речима, приказати однос продатих маркираних и немаркираних производа.
8. Приказати информације о производима који су више пута виђени, а ниједном додати у садржај дигиталне корпе.
9. Приказати пет најатрактивнијих производа у периоду новогодишњих празника.
10. За сваког корисника исписати број прегледа пре прве куповине. Уколико корисник није извршио ниједну куповину, исписати одговарајућу поруку.

4. Обрада података у реалном времену

Прикупљање и обрада података у реалном времену омогућава праћење активних трендова на тржишту. Анализа ових података може бити корисна приликом доношења одлука о прилагођавању пословних стратегија спрам стварних дешавања на тржишту. Уколико се ове промене уведу у право време, значајно могу побољшати успешност пословања. У наредним секцијама налази се опис скупа података који ће бити коришћен за обраду података у реалном времену, заједно са неколико питања за која се очекује ефикасан одговор.

4.1 Опис скупа података

Скуп података, приказан на слици 2, преузет је са сајта *Kaggle.com*. Овај скуп података садржи информације о онлајн поруџбинама на територији Сједињених Америчких Држава. Конкретно, реч је о поруџбинама на велико. Иако су ово такође историјски подаци, модификацијом вредности колоне која уводи временску димензију симулирано је пристизање података у реалном времену.

Order ID	Product	Quantity Ordered	Price Each	Order Date	Customer Shipping Address	City	Store	Category	Customer Gender	Customer Age Range	Discount
236670	Wired Headphones	16	11.99	2019-08-31 22:21:00	359 Spruce St, Seattle, WA 98101	Dallas		Headphones	Male	18-20	0,18
236671	Bose SoundSport Headphones	9	99.99	2019-08-15 15:11:00	492 Ridge St, Dallas, TX 75001	Los Angeles		Headphones	Male	21-25	0,21
236672	iPhone	8	700	2019-08-06 14:40:00	149 7th St, Portland, OR 97035	New York City		Phone	Male	26-30	0,05
236673	AA Batteries (4-pack)	12	3.84	2019-08-29 20:59:00	631 2nd St, Los Angeles, CA 90001	San Francisco		Batteries	Female	31-40	0,08
236674	AA Batteries (4-pack)	16	3.84	2019-08-15 19:53:00	736 14th St, New York City, NY 10001	Boston		Batteries	Female	41-50	0,14
236675	Wired Headphones	16	11.99	2019-08-02 23:54:00	470 Hill St, San Francisco, CA 94016	Dallas		Headphones	Female		50,22
236676	34in Ultrawide Monitor	5	379.99	2019-08-04 19:52:00	470 Cherry St, Los Angeles, CA 90001	Los Angeles		Monitor	Female	18-20	0,17
236677	20in Monitor	3	109.99	2019-08-13 7:16:00	918 6th St, San Francisco, CA 94016	New York City		Monitor	Male	21-25	0,30
236678	Wired Headphones	4	11.99	2019-08-25 20:11:00	58 9th St, San Francisco, CA 94016	San Francisco		Headphones	Male	26-30	0,25
236679	Macbook Pro Laptop	20	1700	2019-08-07 15:43:00	239 Spruce St, Los Angeles, CA 90001	Boston		Laptop	Male	31-40	0,01
236680	LG Washing Machine	9	600	2019-08-09 19:38:00	967 Willow St, San Francisco, CA 94016	Dallas		Machine	Female	41-50	0,04
236681	AA Batteries (4-pack)	9	3.84	2019-08-26 20:52:00	295 1st St, Boston, MA 02215	Los Angeles		Batteries	Female		50,27
236682	AA Batteries (4-pack)	12	3.84	2019-08-19 12:40:00	118 Johnson St, Portland, OR 97035	New York City		Batteries	Female	18-20	0,01
236683	27in FHD Monitor	10	149.99	2019-08-31 15:47:00	196 West St, Dallas, TX 75001	San Francisco		Monitor	Female	21-25	0,23
236684	Lightning Charging Cable	3	14.95	2019-08-09 16:50:00	669 12th St, New York City, NY 10001	Boston		Cable	Male	26-30	0,22
236685	Apple AirPods Headphones	10	150	2019-08-23 19:29:00	238 Highland St, Atlanta, GA 30301	Dallas		Headphones	Male	31-40	0,23
236686	AAA Batteries (4-pack)	8	2.99	2019-08-15 19:13:00	766 Maple St, Dallas, TX 75001	Los Angeles		Batteries	Male	41-50	0,24
236687	USB-C Charging Cable	2	11.95	2019-08-23 12:54:00	668 Meadow St, New York City, NY 10001	New York City		Cable	Female		50,15
236688	34in Ultrawide Monitor	12	379.99	2019-08-08 16:06:00	821 7th St, Los Angeles, CA 90001	San Francisco		Monitor	Female	18-20	0,00

Слика 2. Део скупа података намењеног за обраду у реалном времену.

4.2 Упити

1. За сваку старосну групу, једном дневно, исписати укупан број поруџбина. Поред ове информације, потребно је приказати и укупну вредност поруџбина за сваку групу посебно. На крају, старосне групе сортирати у опадајућем редоследу.
2. На сваких 10 минута, приказати категорију производа са највећим оствареним приходом.
3. Приказати информације о пет производа са највећим бројем поруџбина у последњих 20 минута.
4. На сваких 30 минута, приказати број поруџбина по градовима, и одредити три најактивнија града.
5. У временским интервалима дужине 20 минута, израчунати просечну вредност попушта за сваку категорију посебно.