

Trường Đại Học Sư Phạm Kỹ Thuật TP HCM

Khoa Công Nghệ Thông Tin



Tiểu luận chuyên ngành

Đề tài: Tìm hiểu kiến trúc Delta và ứng dụng

GVHD: Th.s Quách Đình Hoàng

GVPB: TS. Nguyễn Thiên Bảo

SVTH:

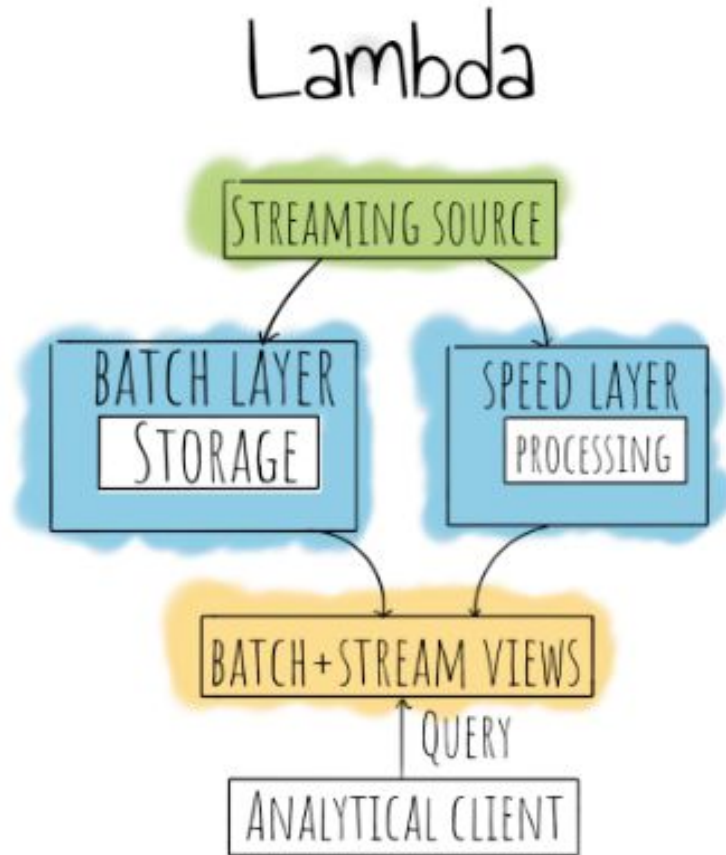
Trần Công Tuấn Mạnh: 19133035

Trần Phát Đạt: 19133018

Nội dung

1. Vấn đề - Những kiến trúc đã có
2. Giải pháp - Kiến trúc Delta
3. Kiến thức nền tảng
4. Ứng dụng

1 - Vấn đề - Kiến trúc Lambda



1 - Vấn đề - Kiến trúc Lambda

Dưới đây là một số ưu điểm của kiến trúc Lambda:

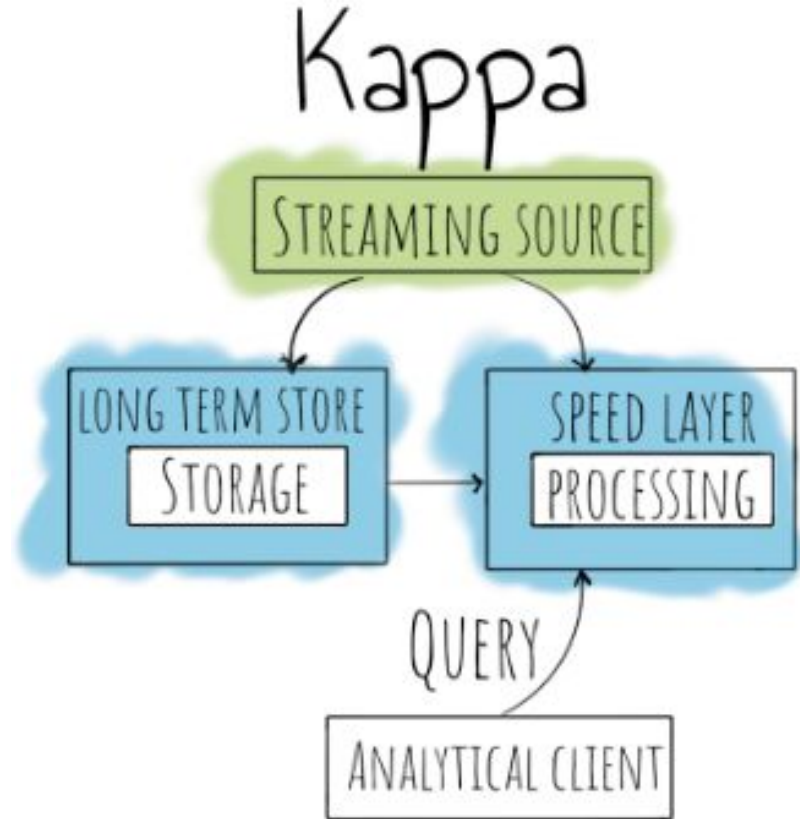
- *Khả năng mở rộng*: Kiến trúc Lambda được thiết kế để xử lý khối lượng dữ liệu lớn và mở rộng quy mô theo chiều ngang để đáp ứng nhu cầu của doanh nghiệp.
- *Khả năng chịu lỗi*: Kiến trúc Lambda được thiết kế để có khả năng chịu lỗi, với nhiều lớp và hệ thống hoạt động cùng nhau để đảm bảo dữ liệu được xử lý và lưu trữ một cách đáng tin cậy.
- *Tính linh hoạt*: Kiến trúc Lambda linh hoạt và có thể xử lý nhiều loại khối lượng công việc xử lý dữ liệu, từ xử lý hàng loạt lịch sử đến kiến trúc phát trực tuyến.

1 - Vấn đề - Kiến trúc Lambda

Dưới đây là một số nhược điểm của việc sử dụng hệ thống kiến trúc Lambda:

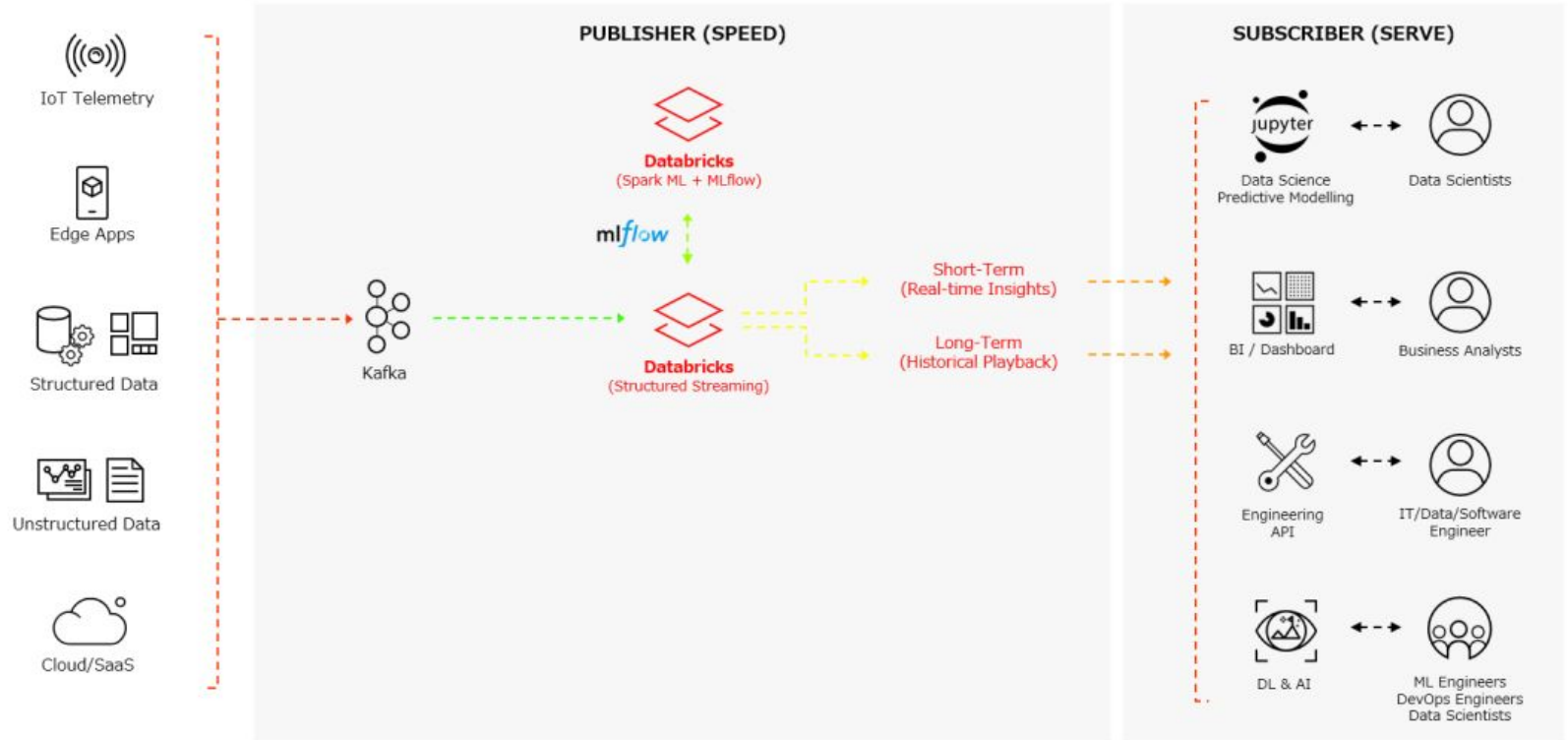
- Độ phức tạp: Kiến trúc Lambda là một hệ thống phức tạp sử dụng nhiều lớp và hệ thống để xử lý và lưu trữ dữ liệu. Mặc dù các lớp của nó được thiết kế cho các đường ống khác nhau, logic có các phần trùng lặp gây ra chi phí viết mã không cần thiết cho các lập trình viên.
- Lỗi và sự khác biệt về dữ liệu: Với việc triển khai gấp đôi các quy trình công việc khác nhau (mặc dù tuân theo cùng một logic, vấn đề triển khai), có thể gặp phải sự cố về các kết quả khác nhau từ các công cụ xử lý hàng loạt và luồng. Khó tìm, khó gỡ lỗi.
- Khóa kiến trúc: Có thể rất khó để sắp xếp lại hoặc di chuyển dữ liệu hiện có được lưu trữ trong kiến trúc Lambda.

1 - Vấn đề - Kiến trúc Kappa



1 - Vấn đề - Kiến trúc Kappa

Kappa Architecture



1 - Vấn đề - Kiến trúc Kappa

Trong kiến trúc Kappa, không cần lớp lô riêng biệt vì tất cả dữ liệu được xử lý bằng hệ thống phát trực tuyến chỉ trong lớp speed. Ngoài những thứ khác, hệ thống như vậy phải có thông lượng xử lý cao và khả năng mở rộng mạnh mẽ để duy trì luồng dữ liệu liên tục không thay đổi.

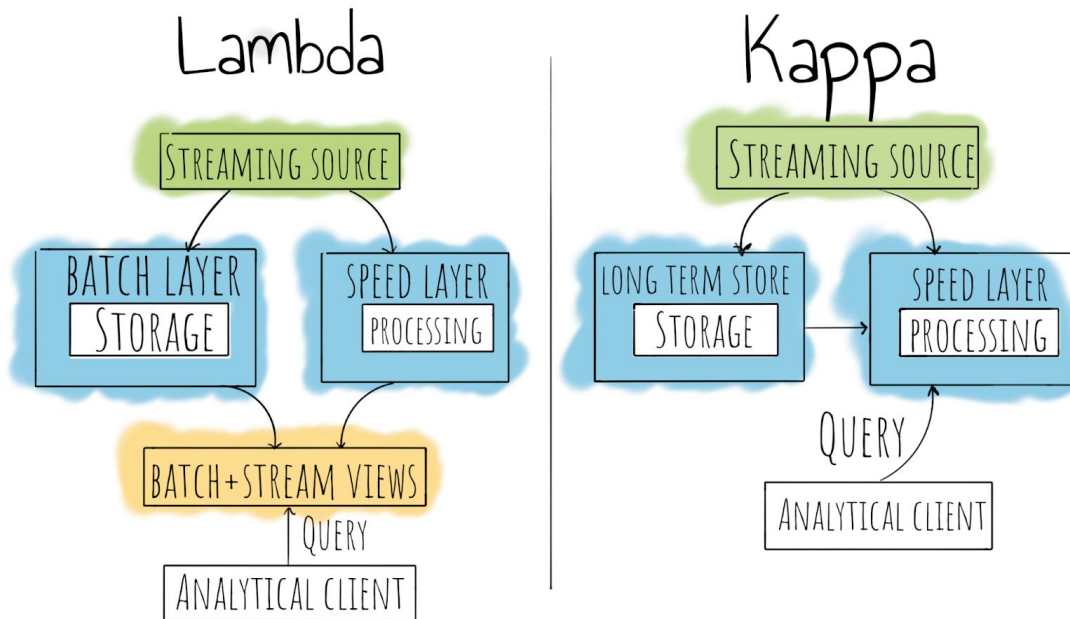
Thay vì xử lý dữ liệu hai lần như trong kiến trúc Lambda, Kappa chỉ xử lý dữ liệu luồng một lần và hiển thị dưới dạng chế độ xem thời gian thực bằng cách sử dụng các công nghệ như Spark.

Tất cả dữ liệu, bất kể nguồn và loại dữ liệu, đều được lưu giữ trong một luồng và người đăng ký (tức là người dùng cuối) sẽ phát lại các luồng được tính toán trước trong khoảng thời gian mong muốn dựa trên trường hợp sử dụng.

Do thiếu kho dữ liệu hợp nhất tất cả dữ liệu tại một nơi, hệ thống phát trực tuyến có thể cần duy trì hàng đợi hoạt động trong nhiều năm (tức là thời gian nó tồn tại) để hỗ trợ các trường hợp sử dụng cần phát lại lượng lớn dữ liệu lịch sử (tương đương với lớp lô trong Lambda).

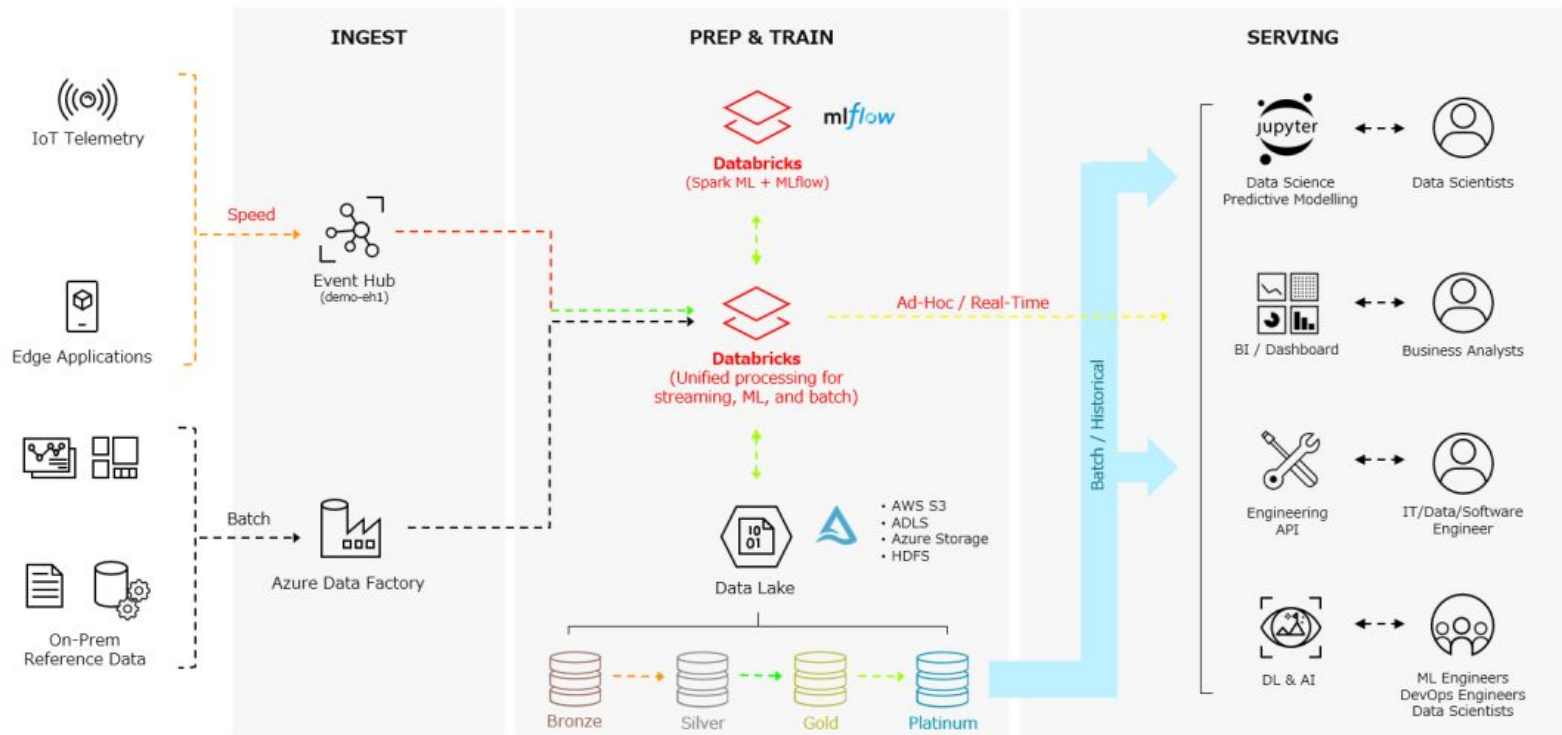
1 - Giải pháp - Kiến trúc Delta

Cho đến gần đây, Lambda và Kappa là hai kiến trúc chính duy nhất để xử lý lượng dữ liệu khổng lồ. Cả hai đều có điểm mạnh và điểm yếu, nhưng chúng ta có thể thấy rằng trong nhiều trường hợp, Lambda là lựa chọn thiết thực hơn do có hồ dữ liệu để xử lý hàng loạt dữ liệu ở trạng thái nghỉ.



2 - Giải pháp - Kiến trúc Delta

Unified Analytics Pipeline Delta Architecture with Databricks



2 - Giải pháp - Kiến trúc Delta

Trong kiến trúc Delta, batch transform có thể thực hiện DML chẳng hạn như thao tác CRUD trên cấu trúc dữ liệu hiện có trong hồ dữ liệu bằng cách sử dụng công nghệ được gọi là Delta Lake. Trên thực tế, Delta Lake mang đến các khả năng giống như Datawarehouse (giao dịch ACID, DML, công nghệ lập chỉ mục chuyên dụng cho bộ dữ liệu phân tán, v.v.) cho Data Lake cũ, giúp nó hoạt động hiệu quả hơn và đáng tin cậy hơn trong các quy trình xử lý Dữ liệu lớn.

Do đó, nó hợp nhất hai lớp một cách hiệu quả để xử lý liên mạch (nghĩa là cùng một công cụ, cùng một API và cùng một mã cho batch và streaming) với chi phí thấp hơn (tức là tối ưu hóa hiệu suất). Lợi ích là rất lớn vì khả năng này có nghĩa là các tổ chức không còn phải xử lý dữ liệu khác nhau dựa trên tốc độ nhập và phương pháp xử lý.

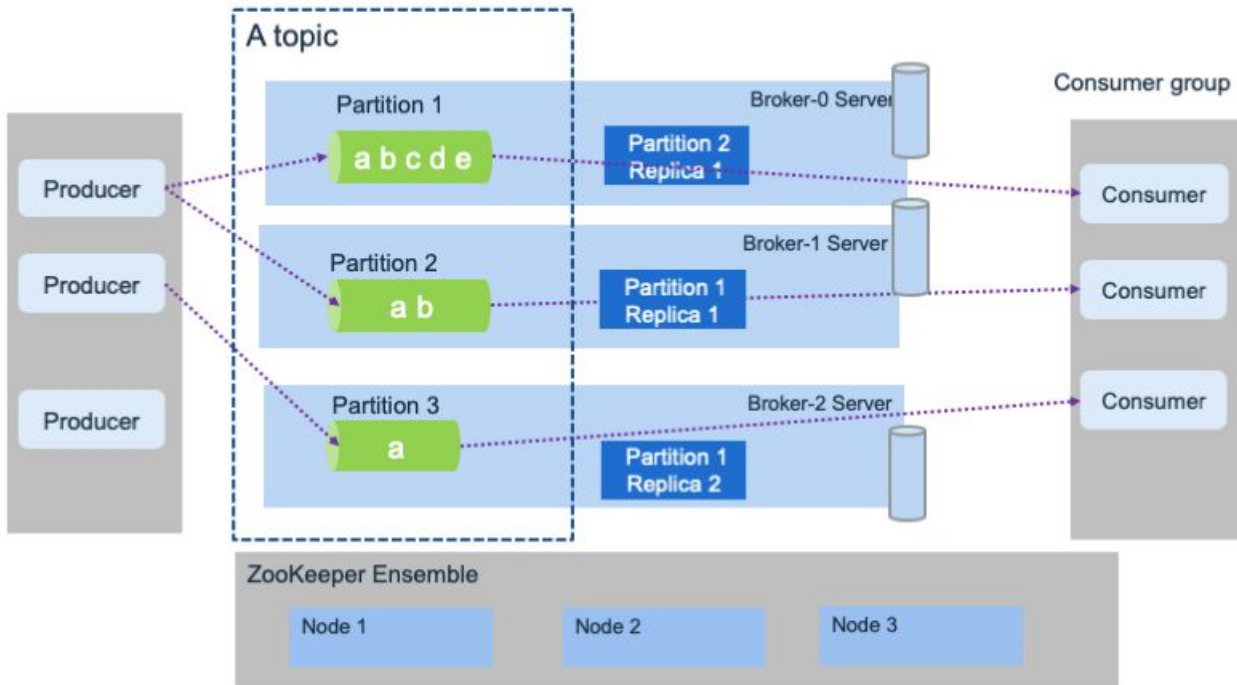
Có thể nhận thấy mẫu này tương tự như Lambda, vẫn giữ được lợi ích của dữ liệu xử lý batch ở trạng thái nghỉ nhưng không cần duy trì một bộ mã riêng cho lớp speed:

Compare

	Lambda	Kappa	Delta
Advantage	append	1 layer	kappa + saving source
Disadvantage	Complex	No Saving Source	Settings

3 - Kiến thức nền tảng - Kafka

Kafka Architecture

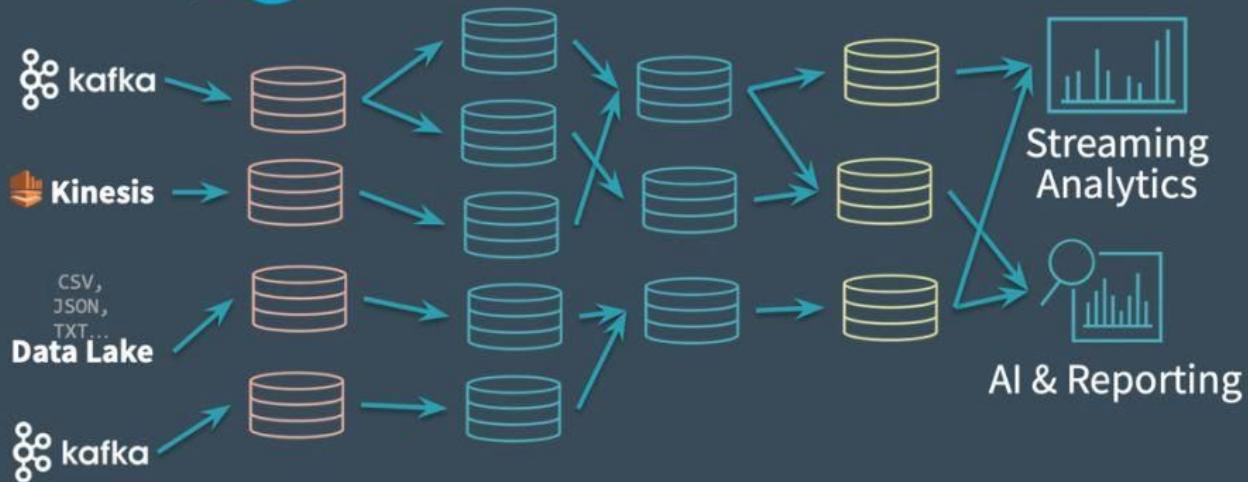


3 - Kiến thức nền tảng - Delta Lake



3 - Kiến thức nền tảng - Delta Lake

The DELTA LAKE Architecture

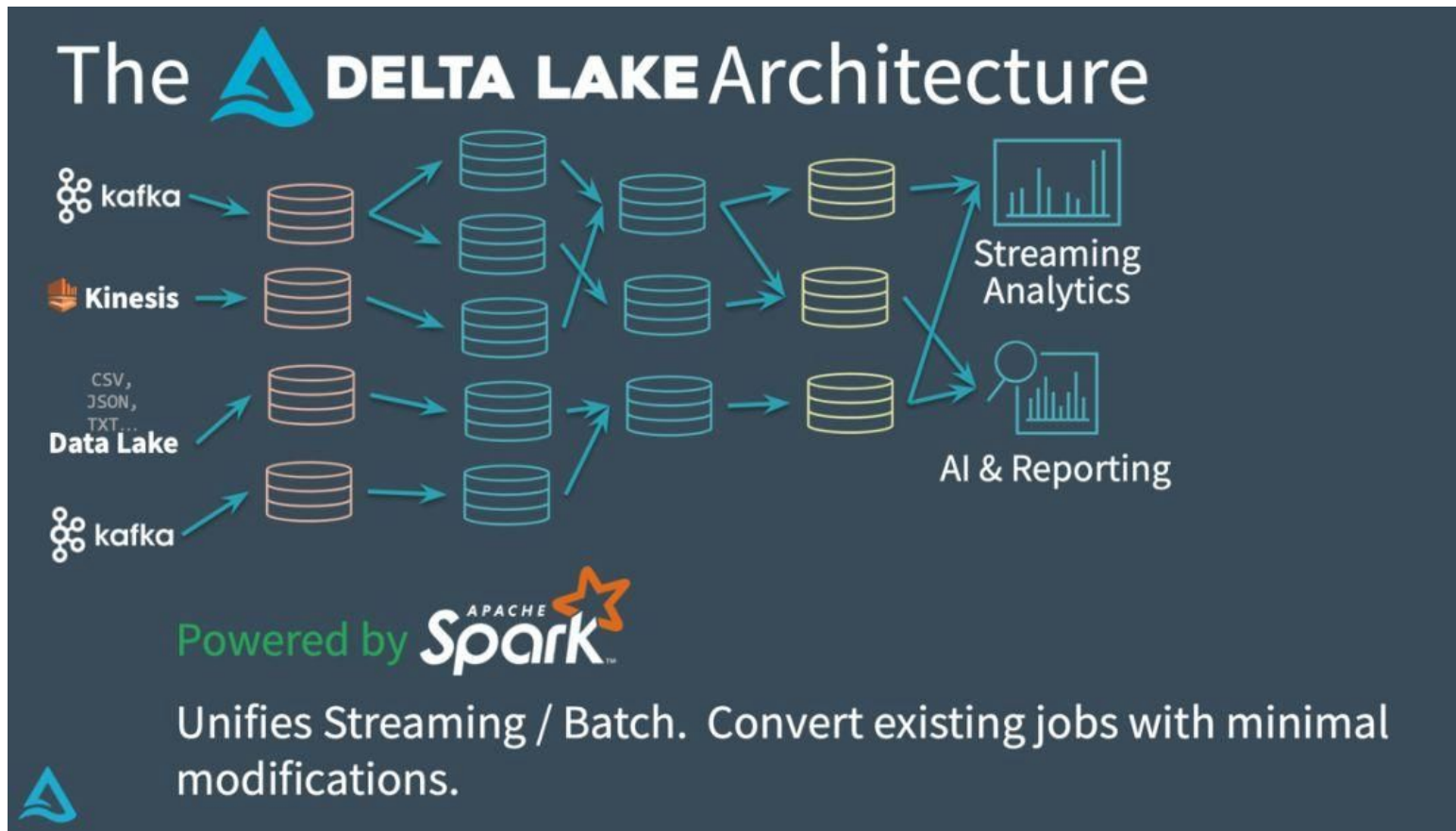


Full ACID Transactions on your Big Data

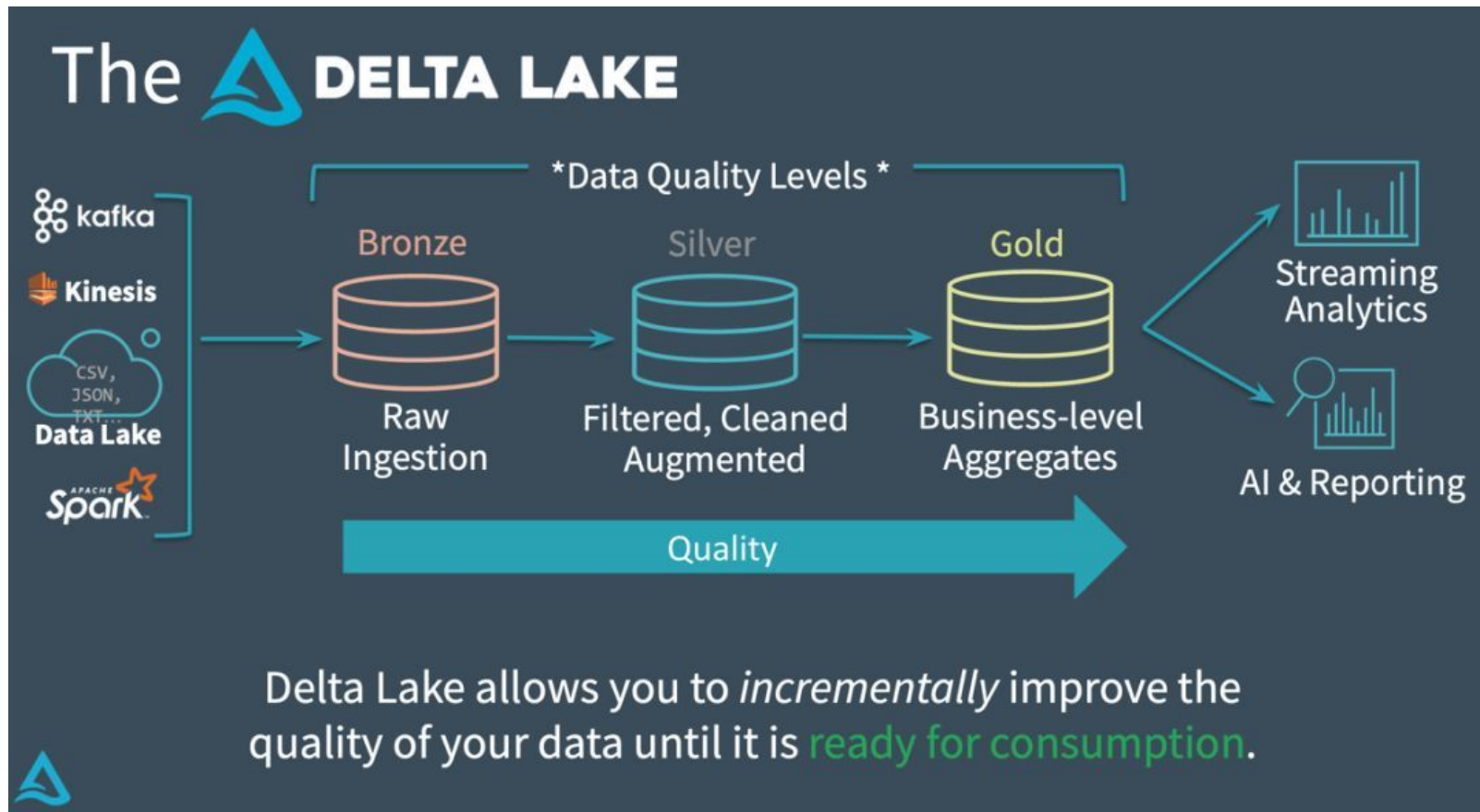
Focus on your data flow, instead of worrying about failures.



3 - Kiến trúc nền tảng - Delta Lake



3 - Kiến thức nền tảng - Delta Lake



3 - Kiến trúc nền tảng - Delta Lake



3 - Kiến trúc nền tảng - Spark Structured Streaming

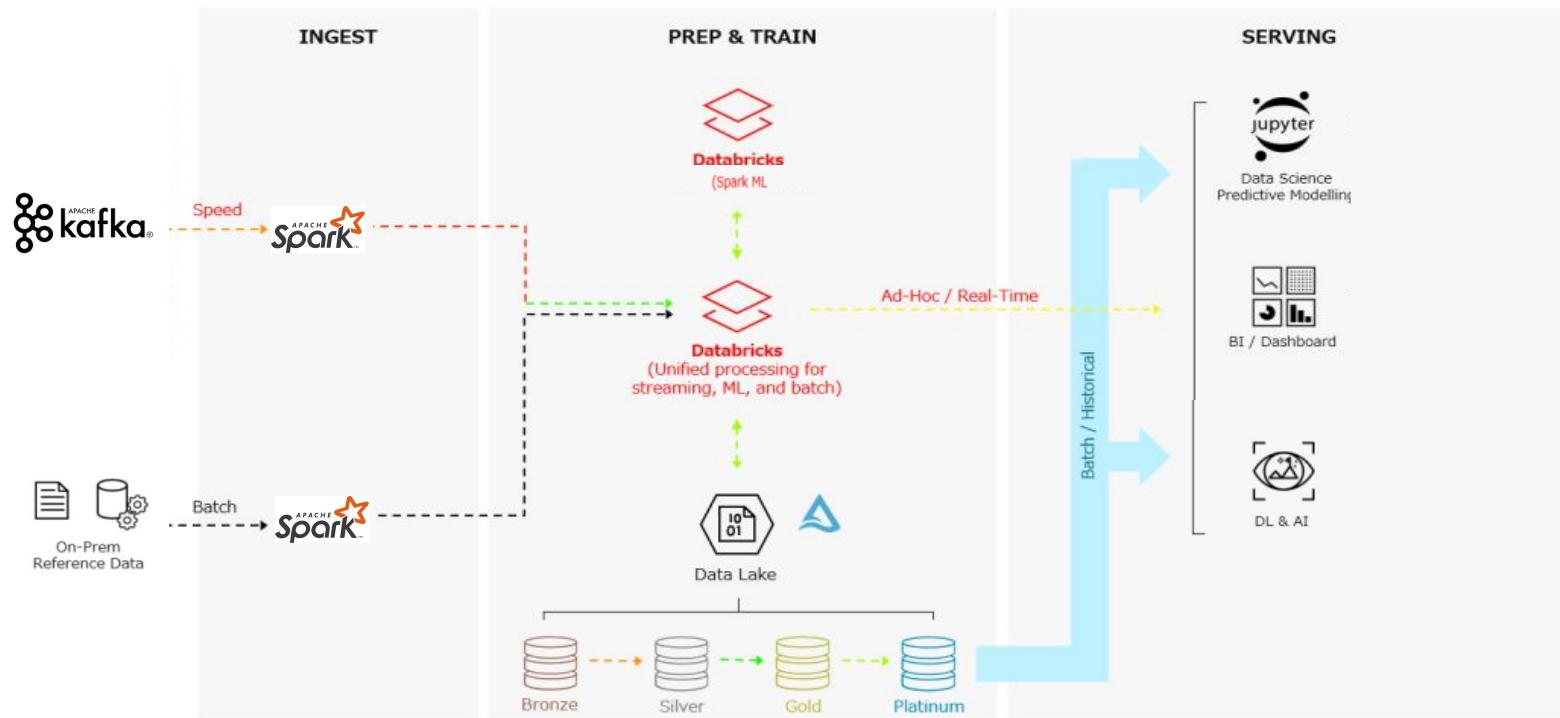


3 - Kiến trúc nền tảng - Delta Lake



4 - Ứng dụng - Kiến trúc

Unified Analytics Pipeline Delta Architecture with Databricks



4 - Ứng dụng - Dữ liệu

This is a restaurant review dataset of each customer's ratings for the restaurants, include information about review, business, checkin, review, tip and user:

Link Data: [Yelp Dataset: Restaurant Ratings](#)

- 6,990,280 reviews
- 1987929 user
- 150,346 businesses

data:

```
|-- yelp_academic_dataset_business.json  
|-- yelp_academic_dataset_checkin.json  
|-- yelp_academic_dataset_review.json  
|-- yelp_academic_dataset_tip.json  
|-- yelp_academic_dataset_user.json
```

4 - Ứng dụng - Dữ liệu

This is ratings data of yelp_academic_dataset_review.json file:

root

```
|-- business_id: string
|-- cool: long
|-- date: string
|-- funny: long
|-- review_id: string
|-- stars: double
|-- text: string
|-- useful: long
|-- user_id: string
```


review_id	user_id	business_id	stars	useful	funny	cool	text	date
WsEr-kocvUQIp111w...	prwt3CRF5IaV_vgI...	SjnbzWbNLAXLOFVIy...	1.0	1	0	0	I dropped my car ...	2018-02-23
aAcQibR3zW0vk4atb...	7P9w2PrP4ZcJyDFwc...	Zi-F-YvyVOK0k5QD7...	5.0	0	0	0	Definitely recomm...	2016-09-18
-up4mW6WdqzGrRh7t...	xbybLiQockAzC4xA1...	EpREWEpmR8f1qLHz...	5.0	0	0	0	After living in t...	2011-11-30
PDHRlnEdkEcwATry4...	UsBxLh14sUp08Sdeq...	Wy8Hswf2cLQGRZN6a...	1.0	1	0	0	If I could give i...	2011-08-24
EHEnA6AIWMSkz44UF...	H-kSAP2ZjKWGZR9Ic...	9dm_v79s9-pefMahq...	4.0	3	0	1	Service was a bit...	2018-07-07
cgppGTg8LpqZ0bn_k...	TSxzAqKzU20vjWSLo...	skN2XhKXlcf53uIw...	5.0	0	0	0	Due to it's locat...	2016-03-07
uIYhNWj3-OsSLH6ef...	5NXzzCTDza-fB00jE...	YGdUUAqeRT5Z7fYkp...	5.0	1	1	0	This review is fo...	2014-11-12
OJ4DSE87REOqg_I8u...	GmSbKlsp0ITmTWO8M...	AWnzFqIr1kLAKTc46...	5.0	0	0	0	This place came h...	2018-03-23
gbK7d1rp0tLKPS8oM...	oMntNOXYFN0qdSCSf...	7sKfrJmjG6unAQeWd...	5.0	0	0	0	I really like Fra...	2016-07-14
vyKiIvMR6aQ5QddXc...	_Bc_E_368qHBi70LA...	aw5GN4yk6r0r9e_5T...	2.0	3	0	0	This use to be a ...	2010-11-22

4 - Ứng dụng - Batch Job - Business Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS business
3 USING DELTA
4 LOCATION '/data/batch/business';
```


batch_business

 business_id	nvarchar
address	nvarchar
attributes	nvarchar
categories	nvarchar
city	nvarchar
hours	datetimeoffset
is_open	nvarchar
latitude	geography
longitude	geography
name	nvarchar
postal_code	nvarchar
stars	int
state	nchar

4 - Ứng dụng - Batch Job - User Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS user
3 USING DELTA
4 LOCATION '/data/batch/user';
```

batch_user	
 user_id	int
average_stars	int
compliment_cool	int
cool	int
elite	int
fans	int
friends	int
funny	int
name	nvarchar
review_count	int
yelping_since	date

4 - Ứng dụng - Batch Job - Checkin Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS checkin
3 USING DELTA
4 LOCATION '/data/batch/check_in';
```


batch_checkin

 business_id	nvarchar
date	date

4 - Ứng dụng - Batch Job - Tip Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS tip
3 USING DELTA
4 LOCATION '/data/batch/tip';
```

tip	
 business_id	nvarchar
compliment_count	varbinary
date	date
text	nvarchar
user_id	nvarchar

4 - Ứng dụng - Streaming Job - Bronze Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS bronze
3 USING DELTA
4 LOCATION '/data/delta/bronze';
```

streaming_bronze

 review_id	nvarchar
user_id	nvarchar
business_id	nvarchar
stars	int
useful	int
funny	int
text	nvarchar
date	date
timestamp	datetimeoffset

4 - Ứng dụng - Streaming Job - Bronze Table

	review_id	user_id	business_id	stars	useful	funny	cool	text	date	timestamp
1	4zopEEpafwm-c_FNpehZYw	JYYYKt6TdVA4ng9lLcXt_g	SZU9c8V2GuREDN5KgyHfJw	5	0	0	0	We were a bit weary about trying the Shellfish Company on the Wharf as more often than not, many places like these (see Cannery Row, Monterey) feast on a captive audience and provide sub-standard fare at high prices. However, emboldened by the perennial good reviews on Yelp, we suppressed our initial observations and went ahead with the trying it out. The place is small, so definitely plan ahead. You will have to wait, so either you know, just do so, or perhaps try to visit outside of peak hour...	2016-05-31	2022-12-27T13:28:05.433+0000
2	IjUhg8tDsUJZ9h0xrwY4Dg	RreNy-tOmXMIten0wi@Og	cPepkJeRMtHapc_b2Oe_dw	4	1	0	1	I was really between 3 and 4 stars for this one. I LOVE the 96th street Naked Tchopstix so I was very excited to see this one which is closer to my house. The vibe is totally different as this is geared more to take out although they do have a decent size dining area. You order at the counter and they deliver it. My daughter and I tried the sushi bowl. You pick up a piece of paper and select from a choice of proteins, vegetables, rice and toppings. I like the fact that it is on paper and you do...	2018-07-17	2022-12-27T13:21:50.816+0000
3	-7LkjSPzfVgnVpuVuRuOow	uAu772Kp5kb-tPfGzmU-lA	2Gyg3li9-m6Z67L_4_BRQ	5	7	0	3	I LOVE Weaver's Way and really disagree with some of the content in previous reviews. WW is not necessarily meant for convenience shopping. I am a single woman and I find that I have no trouble shopping at the co-op, during peak times or otherwise. I actually find the co-op to be a great place to shop if you live alone because I can get smaller quantities of items than if I shop at the corporate stores like Acme. I also enjoy the work hours - I've worked throughout the store, but most recently, ...	2008-12-03	2022-12-27T13:28:15.552+0000
4	-up4mW6WdqzGrRh7t_pLmIA	xybyLQockAzC4xAlzfFrGg	EpREWEpmR8f1qLHzF0AA	5	0	0	0	After living in the STL area for way over 10 years now, I am both ashamed and remorseful to admit that I ate here for the FIRST time just a few weeks ago. I am ashamed because it's a St. Louis tradition and remorseful because I missed out on it for so...many...years! This place is amazing. What is not to love! It's a factory where you can get a tour (Friday-Sunday, noon-5pm on the hour), there's a shop, bar, and dining area. They also host many event with live music on the weekends. It's fun to...	2011-11-30	2022-12-27T13:29:56.842+0000

4 - Ứng dụng - Streaming Job - Silver Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS silver
3 USING DELTA
4 LOCATION '/data/delta/silver';
```

streaming_silver

 review_id	nvarchar
user_id	nvarchar
business_id	nvarchar
stars	int
useful	int
funny	int
cool	int
text	nvarchar
date	date
timestamp	datetimeoffset

4 - Ứng dụng - Streaming Job - Silver Table


	user_id ▲	business_id ▲	stars ▲	date ▲	cool ▲	funny ▲	useful ▲	timestamp ▲
1	eUta8W_HdHMXPzLBBZhL1A	04UD14gamNjLY0IDYVhHJg	1	2015-09-23	1	2	1	2022-12-27T13:18:38.211+0000
2	yfFzsLmaWF2d4Sr0UNbBgg	LHSTtnW3YHCeUkRDGyJOyw	5	2015-08-07	0	0	2	2022-12-27T13:18:58.472+0000
3	wSTuiTk-sKNdcFyprzZAjg	B5XSoSG3SfvQGtKEGQ1tSQ	3	2016-03-30	0	1	1	2022-12-27T13:19:08.618+0000
4	r3zeYsv1XFBRA4dJpL78cw	gmjsEdUsKpj9Xxu6pdjH0g	5	2015-01-03	0	2	0	2022-12-27T13:18:48.336+0000

4 - Ứng dụng - Streaming Job - Gold Table



```
1 %sql
2 CREATE TABLE IF NOT EXISTS gold
3 USING DELTA
4 LOCATION '/data/delta/gold';
```

streaming_gold

 business_id	nvarchar
stars	int
date	date
cool	int
funny	int
useful	int
timestamp	datetimeoffset
year	int


4 - Ứng dụng - Streaming Job - Gold Table

	user_id ▲	business_id ▲	stars ▲	date ▲	cool ▲	funny ▲	useful ▲	timestamp ▲	year ▲
1	4mbLmbA-thaDIZTlgxsaCQ	q0Fi4n7shUTmIxl-mMPVXA	2	2016-07-04	0	0	1	2022-12-15T18:41:04.948+0000	2016
2	qWSAH4MzFbHV6UsseJVIZg	dUctvEfHQccW_uxtRup2QQ	1	2015-12-20	0	0	0	2022-12-15T18:41:53.441+0000	2015
3	0VMuCPgwZlilnxGWfJnxKQ	25Uww0C0wvF9CZ_3B6vWtA	5	2016-07-04	0	0	0	2022-12-15T18:41:29.193+0000	2016
4	f10WH1fXhy-68r4AEhAWA	9OG5YkX1g2GReZM0AskizA	4	2016-01-30	0	0	0	2022-12-15T18:41:41.320+0000	2016
5	CNyXcn0c0V5CFmigqqw-Xg	oY5LFo6Yxxf32ePna6mEUQ	5	2014-12-30	0	0	1	2022-12-15T18:42:17.720+0000	2014
6	0XmgOinrZWNO15DlimRQeg	PpZqIVUAP2i8_hNfpQyKhg	5	2012-03-01	0	0	0	2022-12-15T18:42:29.841+0000	2012

4 - Ứng dụng - Streaming Job - Platinum Table

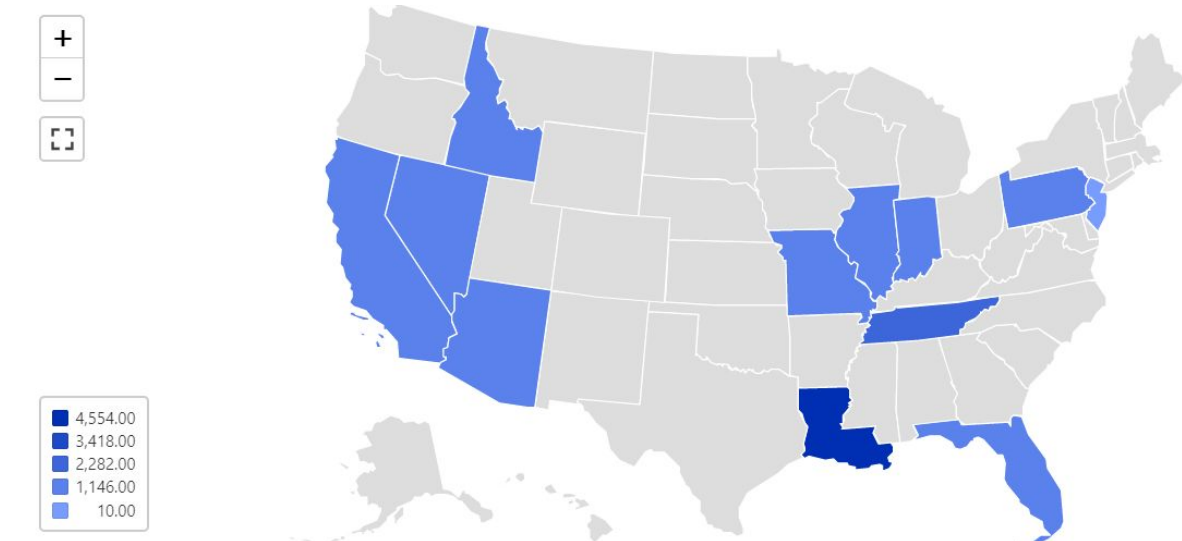


```
1 %sql
2 CREATE TABLE IF NOT EXISTS platinum
3 USING DELTA
4 LOCATION '/data/delta/platinum';
```

streaming_platinum	
 user_id	nvarchar
business_id	nvarchar
stars	int
date	int
cool	int
funny	int
useful	int
year	int
address	nvarchar
categories	nvarchar
city	nvarchar
hours	time
is_open	bit
latitude	geography
longitude	geography
business_name	nvarchar
review_bs_count	bigint
state	nvarchar
user_name	nvarchar
review_count	bigint
yelping_since	time
timestamp	datetimeoffset

4 - Ứng dụng - Streaming Job - Platinum Table

#	user_id	business_id	stars	date	cool	funny	useful	year	address	categories	city	hours	is open	latitude	longitude	business name	review_l	sta	user nai	review	yelping since	timestamp
1	Zp1o52IR92dunms1T96g	gGycnAlpFka_czpO...	5.00	11/04/2013	0.00	0.00	0.00	2013	1 Citizen...	Active Life, Baseball Fields, Sta...	Philadelphia	null	1.00	39.91	-75.17	Citizens Bank Park	515.00	PA	Shannon	354.00	2011-07-03 20:11:27	27/12/2022 20:35:10.615
2	G0DHlgK5dQzzUPWHVEMw	oB8JuukGRgPVVYBFT...	4.00	05/03/2015	0.00	0.00	0.00	2015	121 S 17E...	American (New), Breakfast & ...	Philadelphia	* [17:0-22:0];0:0-0...	1.00	39.95	-75.17	Square 1682	385.00	PA	Aaron	68.00	2010-10-27 20:57:03	27/12/2022 20:24:53.133
3	mmdf_Fi1h3_uZN5z164A	9gObos1OMo6Ugsa...	5.00	06/09/2016	0.00	0.00	0.00	2016	50 S 16th...	American (New), Lounges, NL...	Philadelphia	* [17:0-22:30];17:0...	0.00	39.95	-75.17	R2L	787.00	PA	Mimi	1.00	2012-07-19 12:20:35	27/12/2022 20:31:27.928
4	iYYSitLGp2CpXfKHmefw	Zx7n8mct8OzRXVz...	5.00	27/04/2018	0.00	0.00	0.00	2018	214 11th...	American (New), Restaurants, ...	Nashville	* [6:30-15:0];0:0-0...	1.00	36.15	-86.78	Milk and Honey Nashville	1725.00	TN	Ryan	12.00	2017-02-11 03:51:15	27/12/2022 20:24:22.764
5	XBXCFMZn8pFIWEZckukZw	Zx7n8mct8OzRXVz...	5.00	21/01/2018	0.00	0.00	0.00	2018	214 11th...	American (New), Restaurants, ...	Nashville	* [6:30-15:0];0:0-0...	1.00	36.15	-86.78	Milk and Honey Nashville	1725.00	TN	Namilla	19.00	2017-09-04 21:26:41	27/12/2022 20:37:12.075
6	EZjT2qjN0mOKypMAqZd5rQ	A2q7d_CBM2_81Vtk...	2.00	08/07/2017	0.00	1.00	1.00	2017	345 N Vir...	American (New), Restaurants, ...	Reno	* [16:0-22:0];0:0-0...	0.00	39.53	-119.81	The Buffet	584.00	NV	Tiana	133.00	2013-11-14 19:22:14	27/12/2022 20:24:43.011
7	zoBajFyVA0z4JbF5Miksg	c_lg56PkvMyak7Rbr...	4.00	08/06/2015	0.00	0.00	0.00	2015	1414 Ma...	American (New), Restaurants, ...	Speedway	* [11:0-0:0];11:0-2...	1.00	39.79	-86.24	Barbecue and Bourbon	284.00	IN	Brad	27.00	2010-08-05 23:31:53	27/12/2022 20:22:00.936
8	OuateND39GZQcm2xKBWUQ	JNL5KUP2_4HJUM...	1.00	24/08/2014	0.00	0.00	1.00	2014	9818 US...	American (New), Restaurants, ...	Port Richey	null	0.00	28.30	-82.71	El Chicanito Mexican Res...	10.00	FL	Jenn	16.00	2013-06-13 21:47:57	27/12/2022 20:25:43.729
9	Qac1yFyCFLmCT26YEEBuA	29HqjwOG6uAWqBIZ...	1.00	29/05/2012	0.00	1.00	0.00	2012	702 Anac...	American (Traditional), Americ...	Santa Barbara	* [11:0-22:0];11:0-...	0.00	34.42	-119.70	Paradise Cafe	290.00	CA	Guthrie	28.00	2011-03-13 16:38:04	27/12/2022 20:39:03.391
10	qVYIGrmY6uemAy_1DpfaA	9OfXjYsG5keaaUM...	5.00	26/05/2014	0.00	0.00	4.00	2014	600 Marr...	American (Traditional), Bars, A...	Nashville	* [6:0-1:0];6:0-1:0...	0.00	36.15	-86.69	Champions	73.00	TN	Adam	171.00	2012-05-06 14:33:46	27/12/2022 20:31:38.047
11	1EecTw_lojld62ySMH5gA	xR3inMR2KzeU3b9d...	4.00	30/01/2016	0.00	0.00	0.00	2016	1324 Nie...	American (Traditional), Restau...	Granite City	* [11:0-23:0];11:0-...	0.00	38.70	-90.15	Lacelles Granite City	46.00	IL	Brandi	7.00	2016-01-25 14:33:58	27/12/2022 20:37:32.316



Yelp Restaurant Recommendation System

Preprocessing

- Convert user_id vs business_id to int (auto increment)
- select columns: userid, businessid, stars

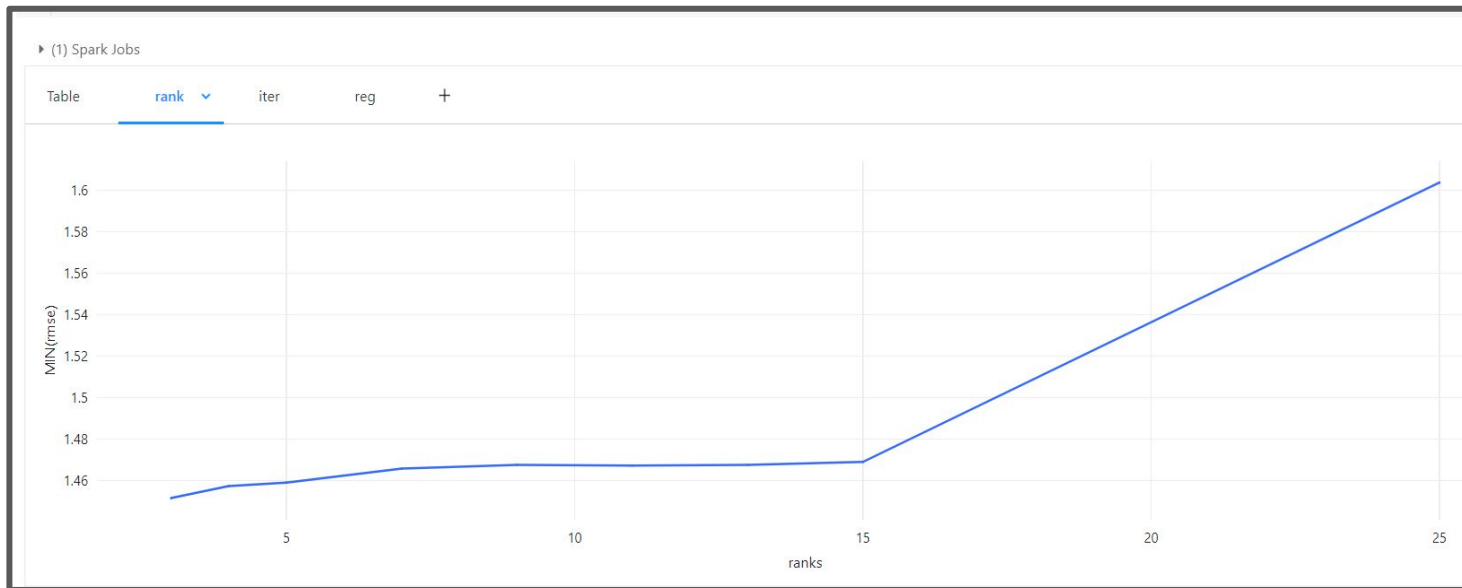
Split data

- Other by date
- Into training, validate, test : 7-2-1/ 10

Yelp Restaurant Recommendation System

Algorithms: ALS (alternating least square)

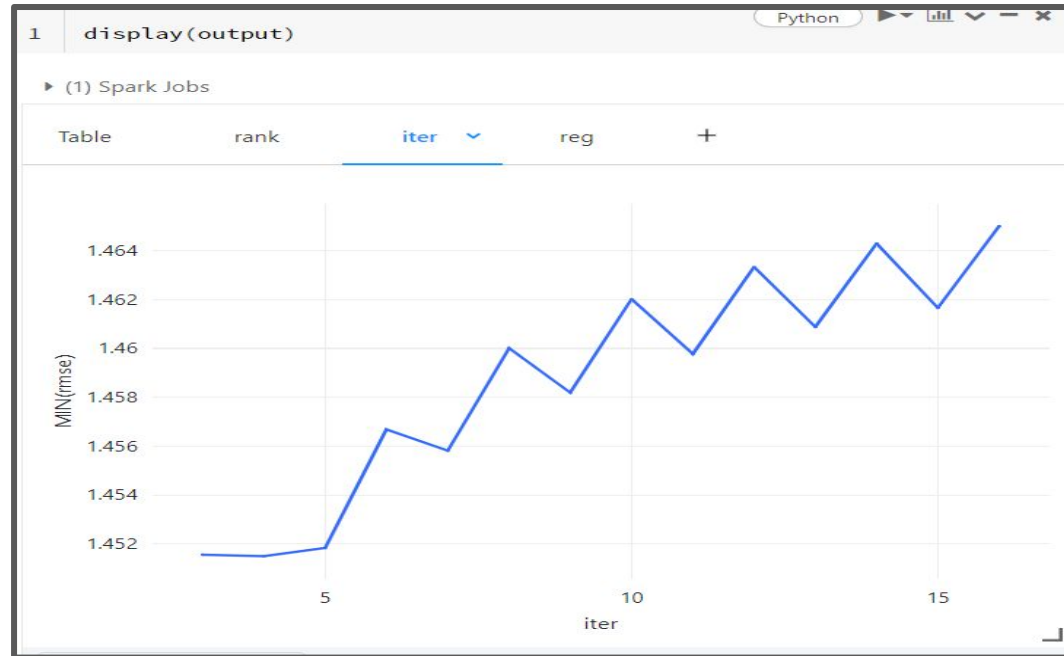
Tuning: rank: => the best is [3,5]=> choose 4



Yelp Restaurant Recommendation System

Algorithms: ALS (alternating least square)

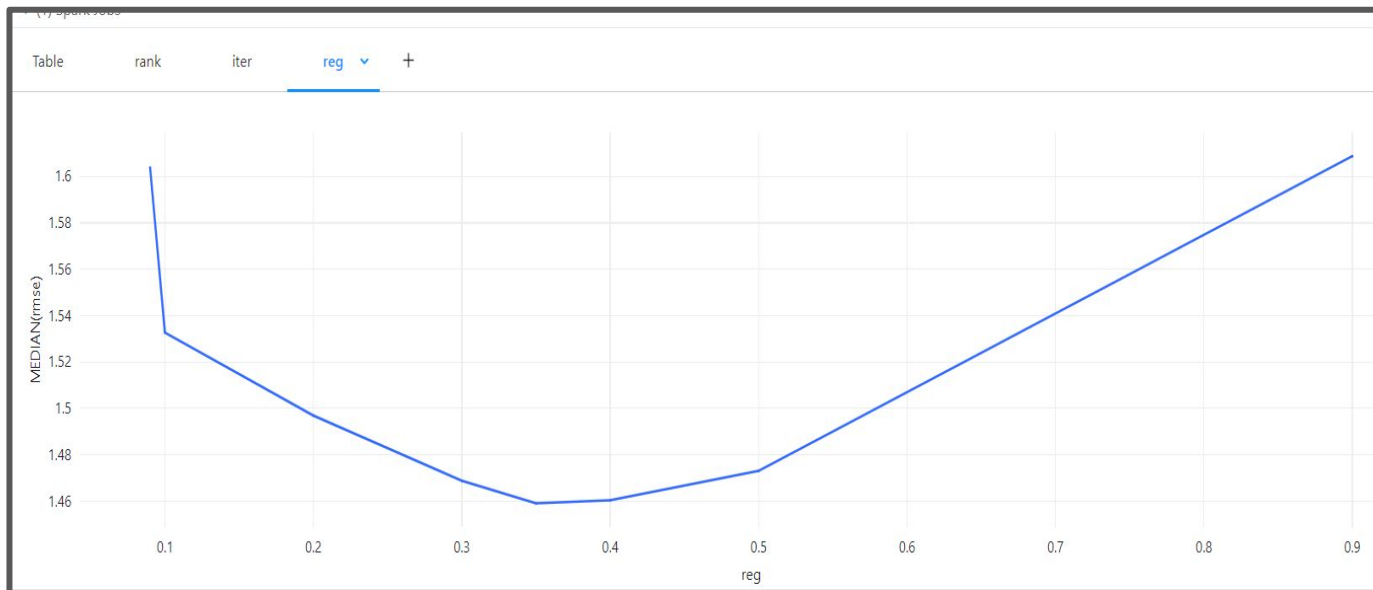
Tuning: maxIter: => the best is around (3,5)=> Choose 4



Yelp Restaurant Recommendation System

Algorithms: ALS (alternating least square)

Tuning: regularization \Rightarrow The best is 0.35



Yelp Restaurant Recommendation System

So We have the the best param: rank: 4, MaxIter: 4, regularization: 0.35

RMSE: 1.3710067457429629

Then we save model:

```
/databricks/driver/yelp_dataset/save  
itemFactors metadata userFactors
```

And use for the specific user: this recommend for user:

userid	businessid	rating
20	26777	5.2499027
20	59308	4.724409
20	13385	4.68757
20	41707	4.6628594
15	26777	5.0053034
15	26624	4.466009
15	41707	4.3750315
15	13385	4.3227687

Conclusion

Have done:

Delta Architecture Knowledge

Demo of Delta Architecture

Basic Application

Remark:

Hard vs Complicated

New

Development:

Build app using delta architecture then recommend for user

THANKS