Enriching the data vs Filtering in Spark

Gokul Prabagaren
Engineering Manager
Card Loyalty,CapitalOne

CapitalOne

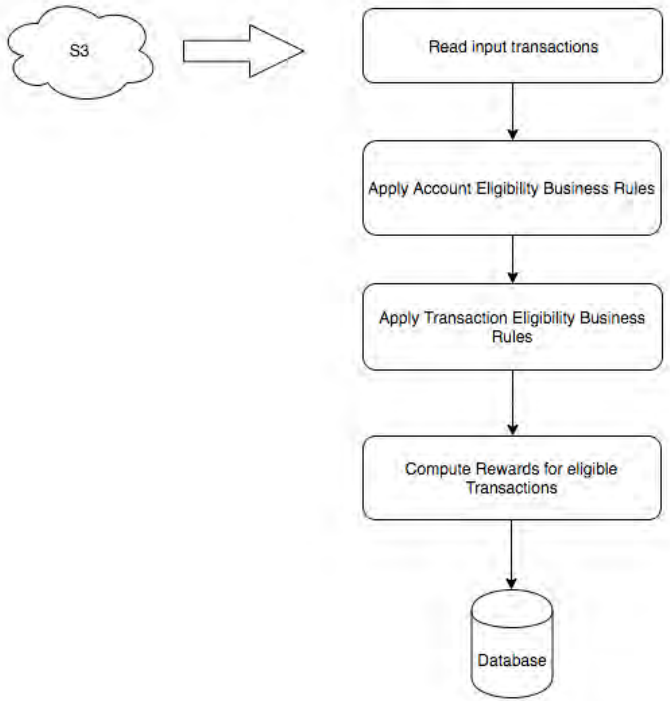
- CapitalOne is the **first U.S Bank** to exit **on-prem legacy data centers** and go all in on the cloud
- CapitalOne is a Tech Company doing **Banking**
- CapitalOne is **Founder-Led** company on mission to change banking for good
- Giving back to the community
  - Open source - 20+ projects like Criticalstack,Rubicon,Dataprofiler
  - CODERS
  - CODA

Agenda

- Loyalty use case in CapitalOne
- Filtering Approach
- Issues with Filtering Approach
- How Enriching approach solves the issue
- Conclusion & Questions

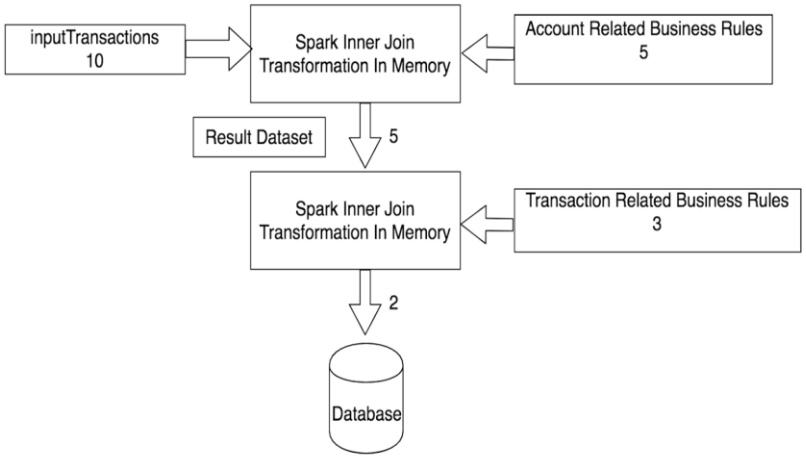
Loyalty Use case in CapitalOne

- Use case
  - One of Core Credit Card Rewards Spark Application.
  - Consumes daily credit card transactions and computes the Rewards

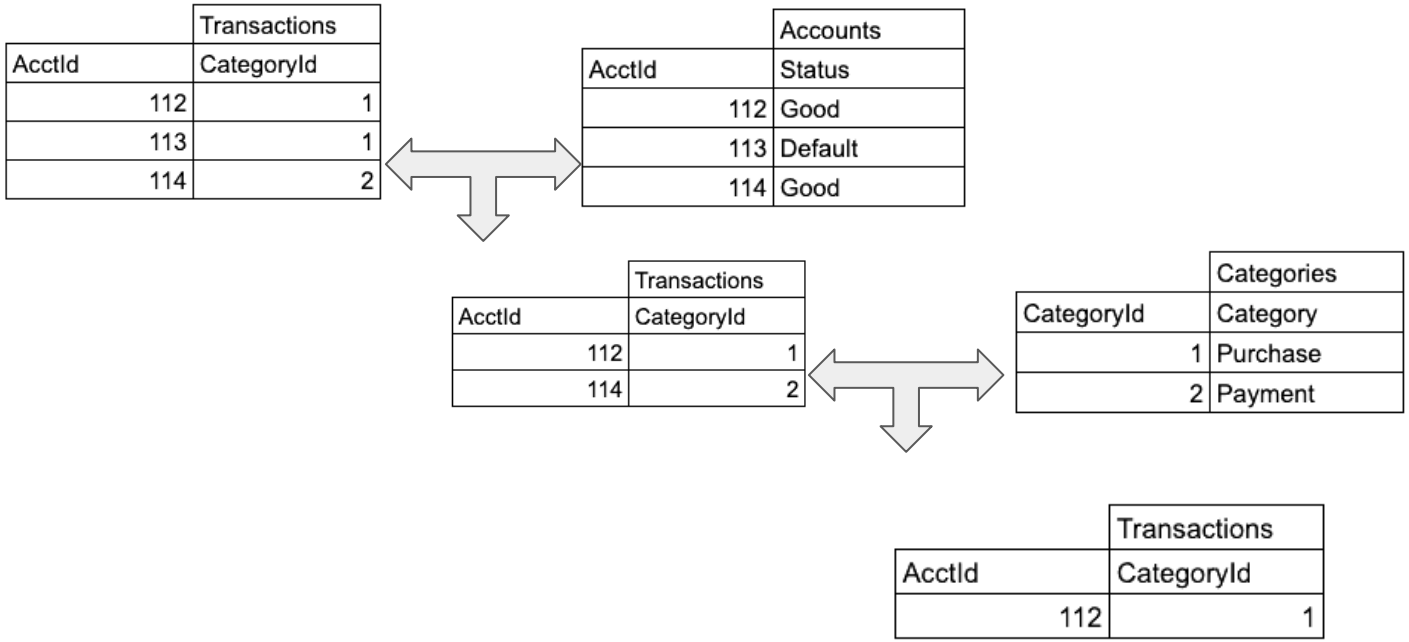


Filtering the data Approach

- This approach uses Spark inner-join at each stage



Filtering Approach Example



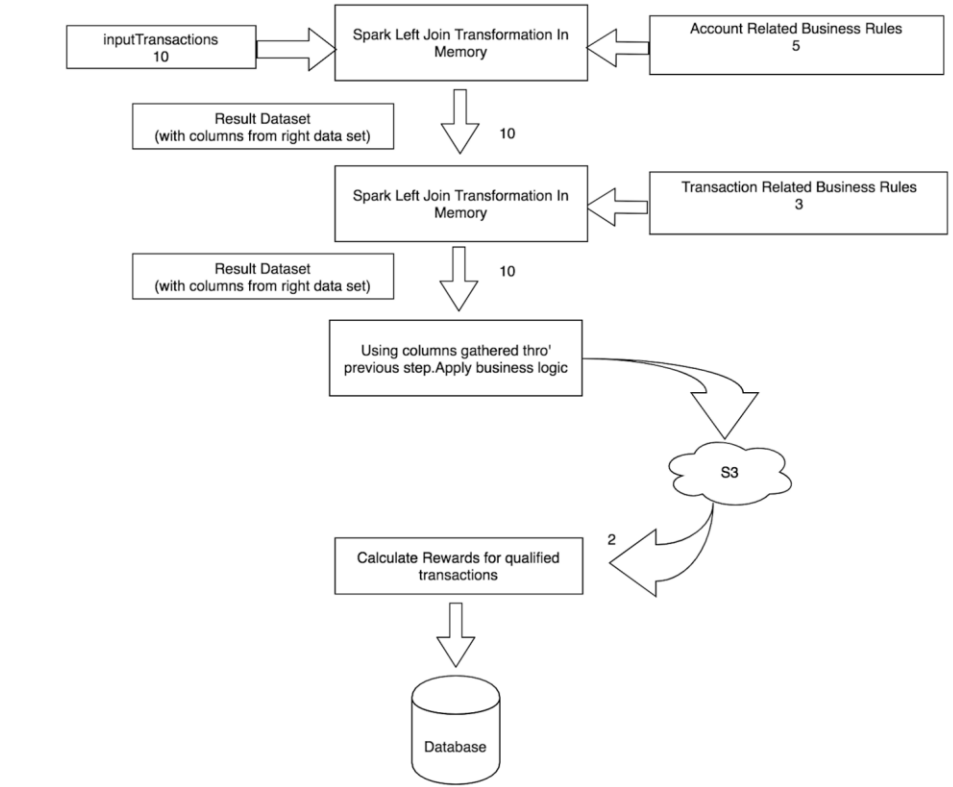
Issues with Filtering Approach

- Hard to debug the application post deployment
- Back tracing of data is not possible as computation happens in-memory
- Counts at each stage can only provide how many got processed.But not why the remaining got dropped in that stage.

How did we overcome these issues ?

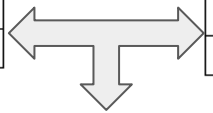
Enriching the data approach

- This approach uses Spark left-outer join
- Instead of filtering the data from dataset at each stage.Enriching approach keeps enriching the data from right side dataset



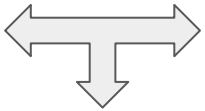
Enriching Approach Example

	Transactions
AcctId	CategoryId
112	1
113	1
114	2



	Accounts
AcctId	Status
112	Good
113	Default
114	Good

	Transactions	
AcctId	CategoryId	Status
112	1	Good
113	1	Default
114	2	Good

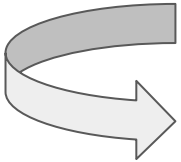


	Categories
CategoryId	Category
1	Purchase
2	Payment

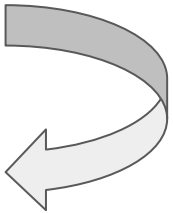
	Transactions		
AcctId	CategoryId	Status	Category
112	1	Good	Purchase
113	1	Default	Purchase
114	2	Good	Payment

Enriching Approach Example…….

	Transactions		
AcctId	CategoryId	Status	Category
112	1	Good	Purchase
113	1	Default	Purchase
114	2	Good	Payment



	Transactions			
AcctId	CategoryId	Status	Category	isEligibleForNextStage
112	1	Good	Purchase	TRUE
113	1	Default	Purchase	FALSE
114	2	Good	Payment	FALSE



	Transactions
AcctId	CategoryId
112	1

Advantage of Enriching over filtering

- Data from each stage is enriched into original dataset.It captures the state information,makes it easy to debug/analyse later

- Same data columns/flags captured at each stage gives more granular details to know why particular data got dropped at that stage

- No need of additional costly counts action at each stage.

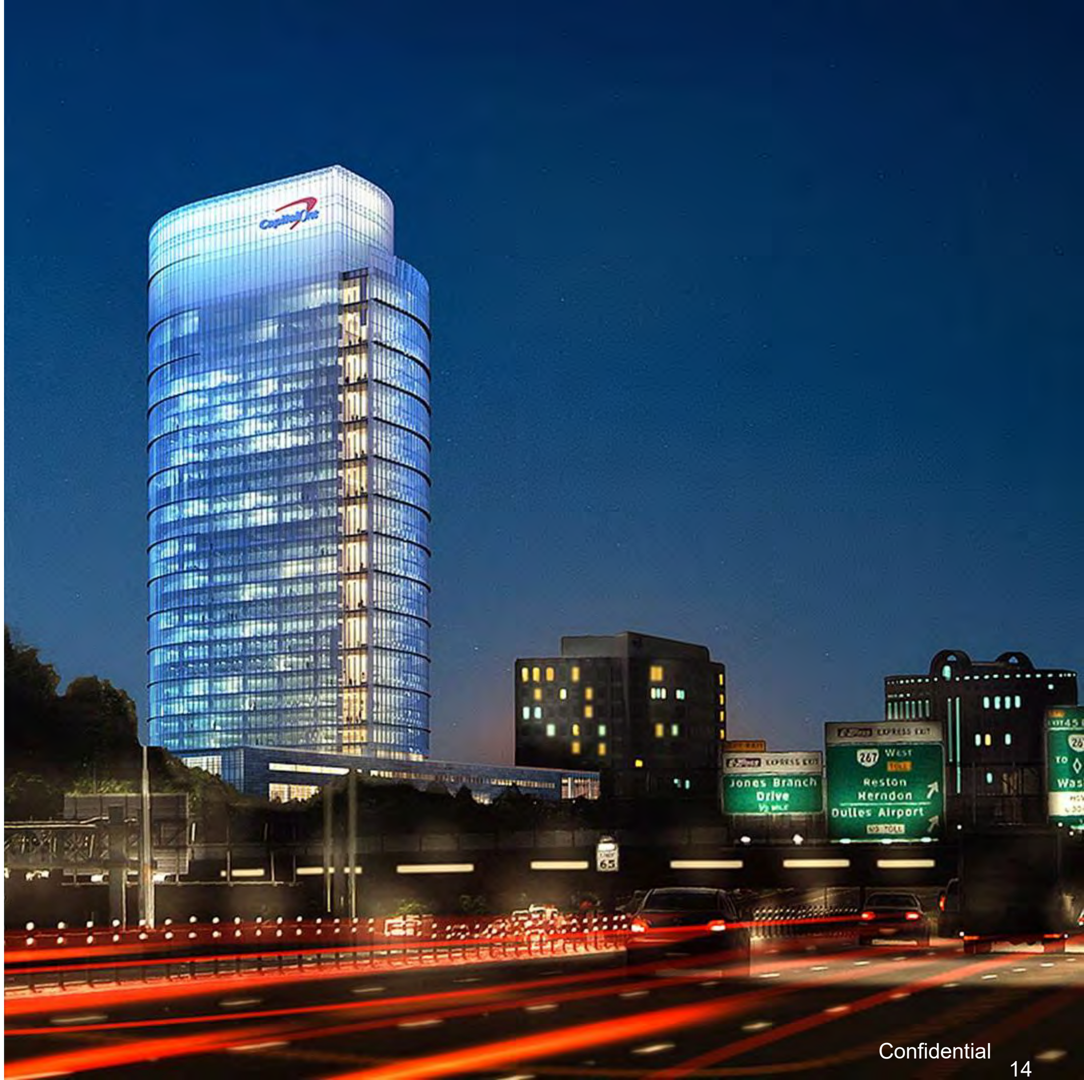
Conclusion

- We made the switch to use Enriching approach in our Spark job in production.
- It is successfully processing millions of credit card transaction daily.
- Awarding millions of miles,cash and points as Rewards to Capital One customers.

About me - Gokul Prabagaren

- Engineering Manager @ Capitalone
- Been building Software Applications since initial versions of Java and Spark
- Tech Speaker and Contributor to @CapitalOneTech Medium blogs on Big Data processing
- @gocoolp on LinkedIn
- @gocool_p on Twitter


Thank You
Vinaka
köszönöm
Kiitos
Maake
Asante
Shukria
Dhanyavadagalu
감사합니다
Dankscheen
спасибо
Kam Sah Hammida
ارکش
Manana
Dankon
Dolch i Chi
Terima Kasih
Matondo
Dank Je
Mauruuru
Biyan
Taiku
Tack
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Arigato
Grazie
Juspaxar
Mochchakkeram
Gracias
Khap
Tingki
நன்றி
Bedankt
Dakujem
धन्यवाद
cảm ơn bạn
Paldies
Ua Tsaug Rau Koj
Grazas
Di Ou Mèsi
Gratias Tibi
Obrigado
Děkuji
Nuwun
Nirringrazzjak
Kia Ora
Kop Khun
ありがとう
Suksama
Rahmat
Hvala
Djiere Dieuf
Welalin
谢谢
хвала
Misaotra
Matur
Danke
Merci
Go Raibh Maith Agat
Eskerrik Asko
Salamat
ขอขอบคุณคุณ
Tuke
Najis


Capital One
Jones Branch Drive
West
Reston
Herndon
Dulles Airport

Capital One

Capital One

What's in your wallet?®