Enriching the data vs Filtering in Spark

Gokul Prabagaren

Confidential

About Me

- Master Software Engineer @ Capitalone
- Been building Software Applications since Java 1.2
- Contributor of @CapitalOneTech Medium blogs on Big Data processing
- @gocoolp on
- @gocool_p on

Agenda

1. Quick Intro on Apache Spark

2. Context of use case in CapitalOne

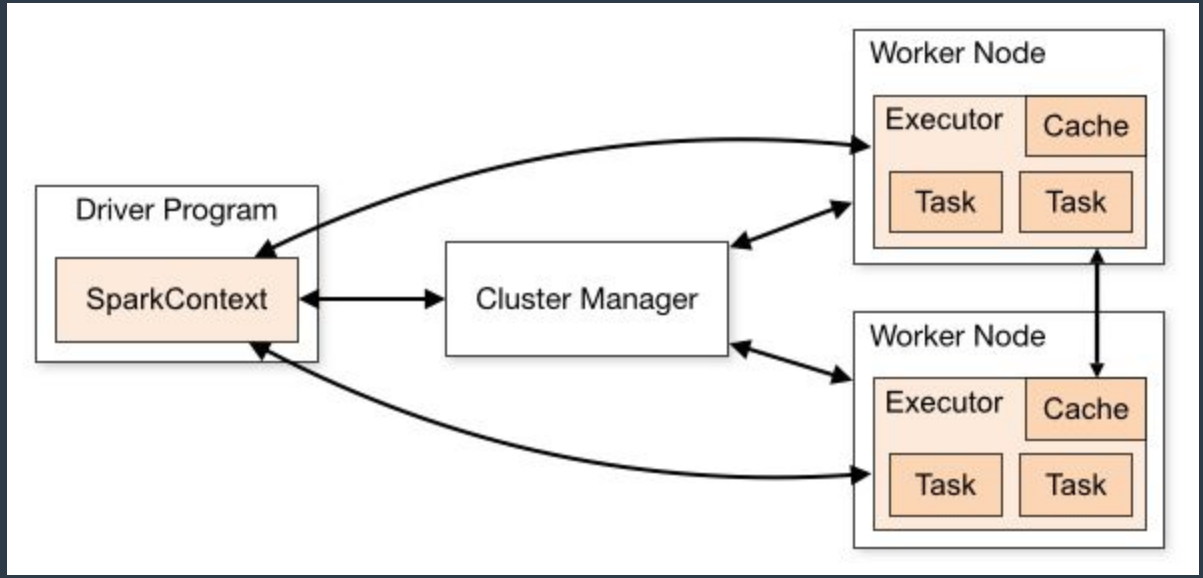
3. Filtering Approach

4. Issues with Filtering Approach

5. How Enriching approach solves the issue

6. Conclusion & Questions

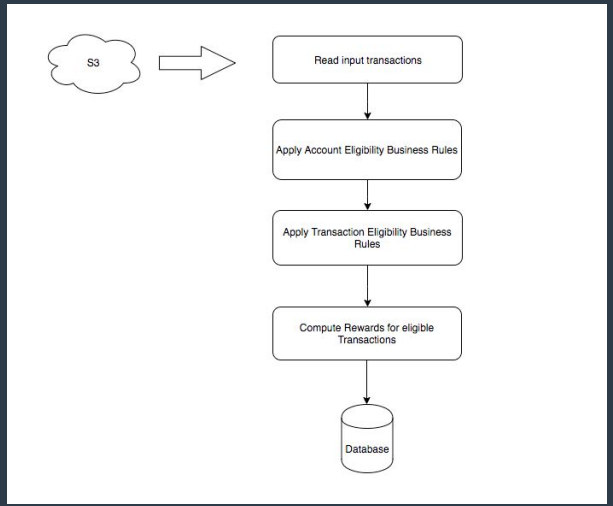
Apache Spark



Source : https://spark.apache.org/docs/latest/cluster-overview.html

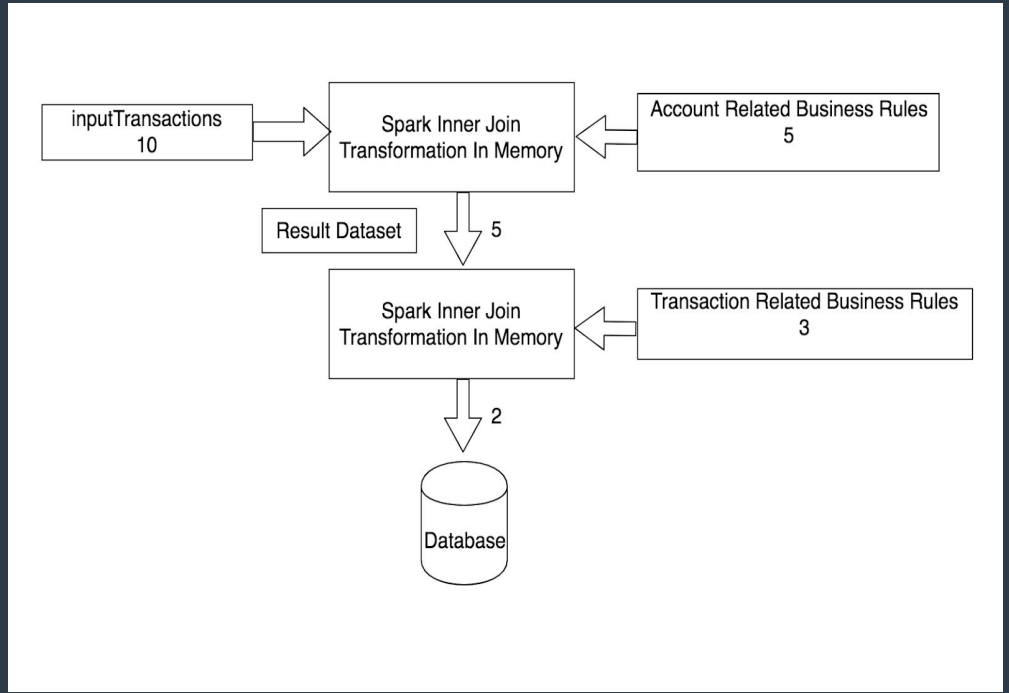
Capital One's Rewards Use case

- CapitalOne develops its software as Open source first in cloud.
- We use Apache Spark extensively for variety of batch,streaming and machine-learning workloads.
- 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



Issues with Filtering Approach

- Hard to debug the application post deployment

- Back tracing of data is not possible as computation happens in-memory

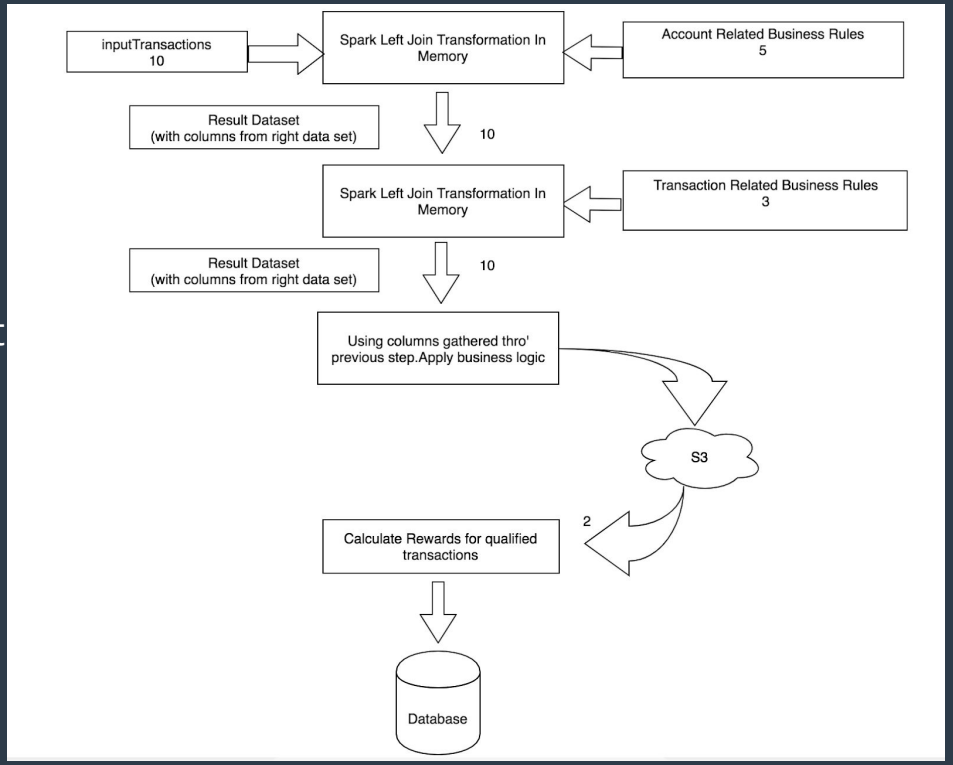
- Counts at each stage can only provide how many got processed.But not why the remaining got dropped in that stage.

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



Advantage of Enriching over filtering

- Since the data from each stage is enriched into original dataset.It captures the state information make 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.

We are hiring

https://www.capitalonecareers.com/

Capital One

What's in your wallet?®