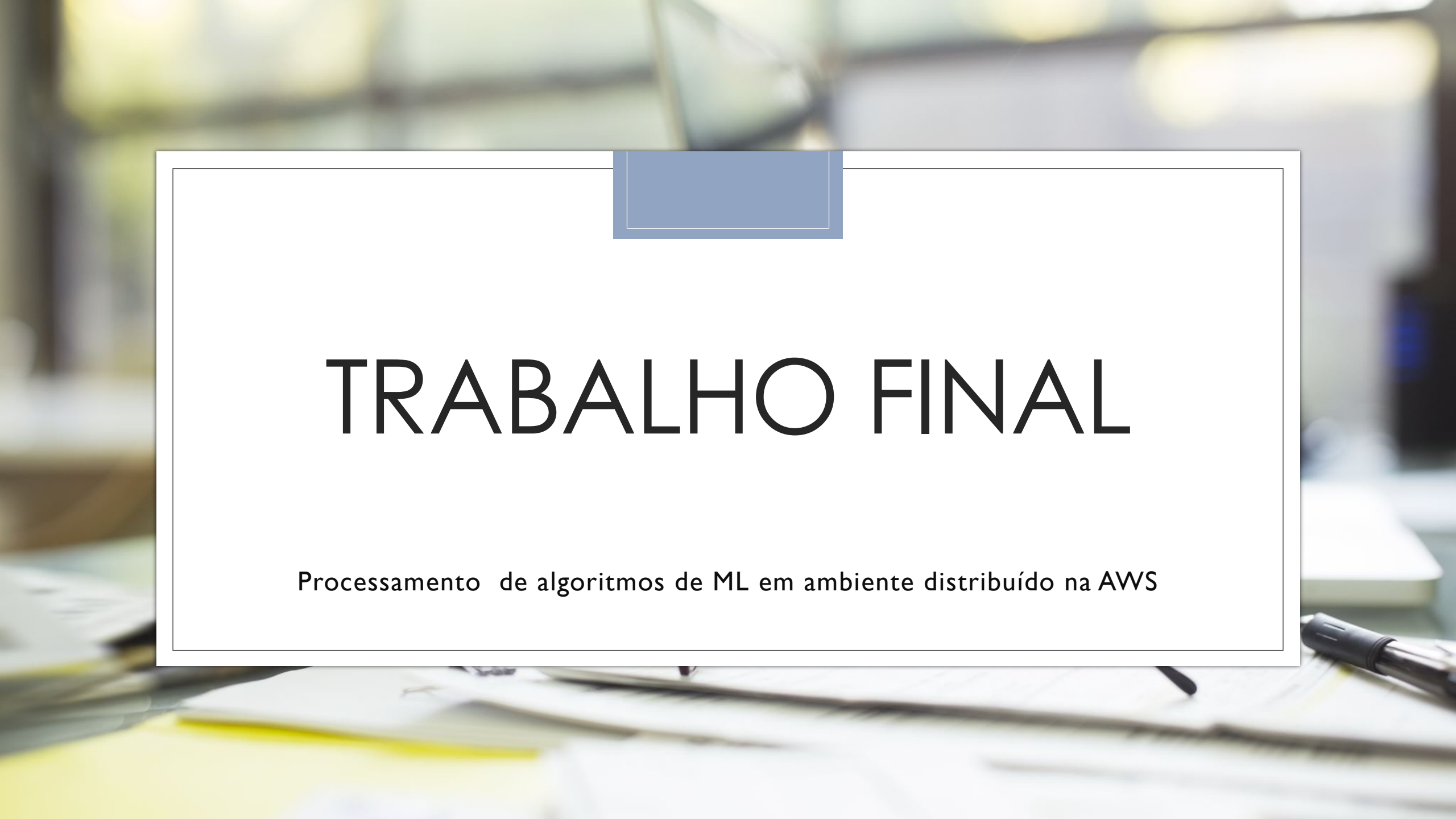


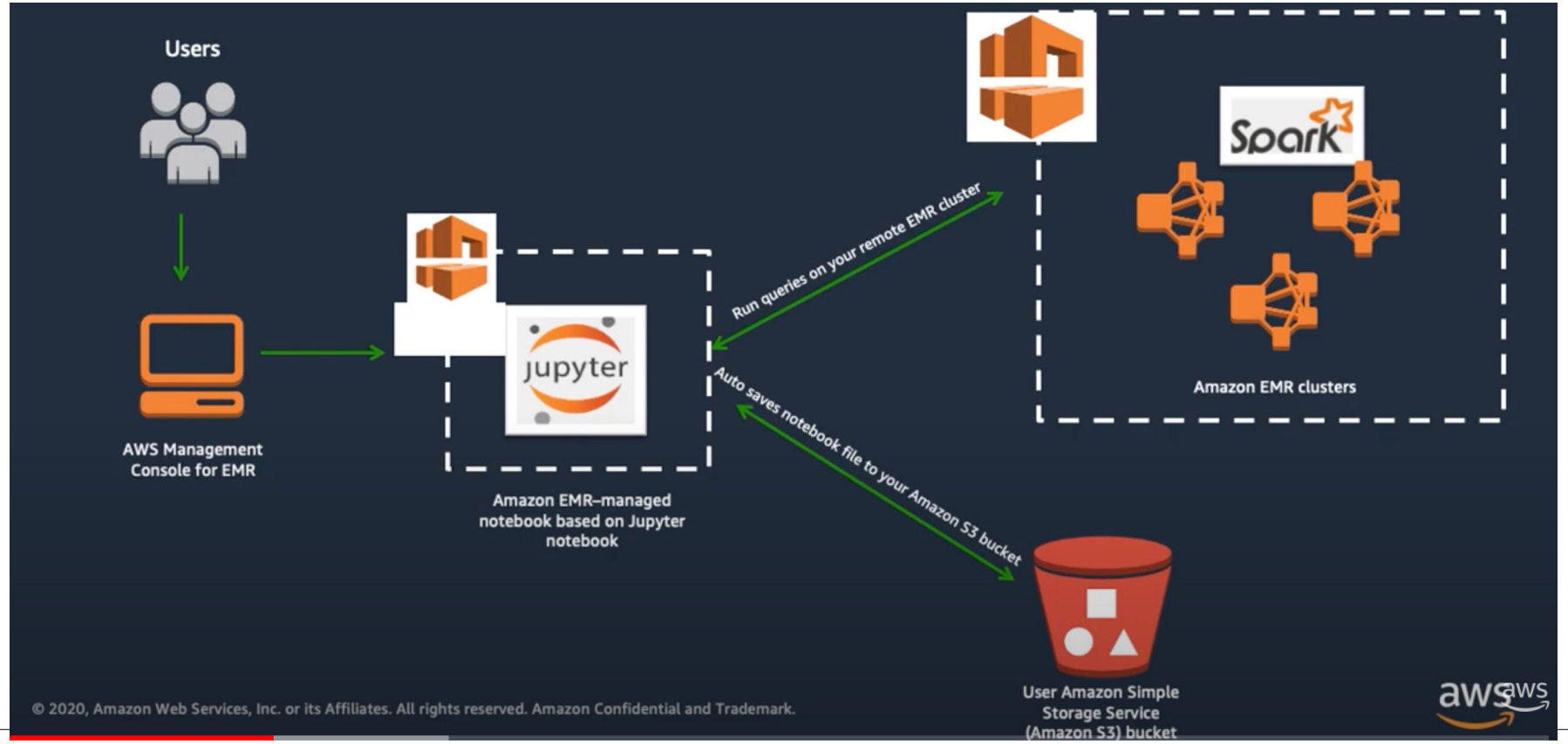
TRABALHO FINAL

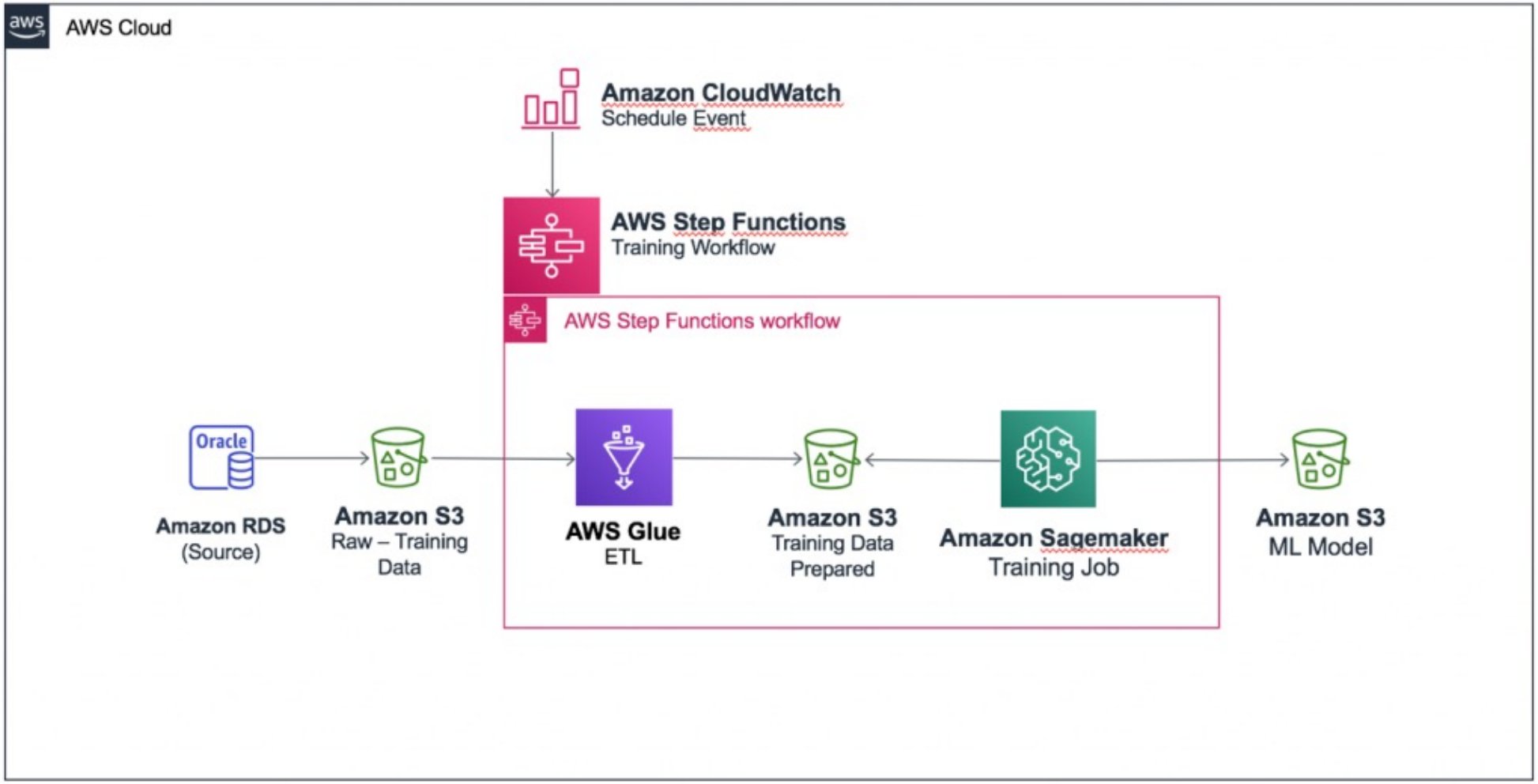
Processamento de algoritmos de ML em ambiente distribuído na AWS

Organização do trabalho

1. Introdução, que deve ser uma breve contextualização.
2. Descrição dos dados (fonte, período, volume, formato, amostras dos dados)
3. Workflow/pipeline de arquitetura
4. Infraestrutura (Configuração do cluster como nro de nós, tipo de máquinas, permisões)
5. Análise do Processamento (análise do desempenho das operações ETL, segundo critérios como tempo de processamento, formas intermediárias de armazenaneto dos dados (partições), comparação do desempenho com diferentes configurações de cluster)
6. Funções definidas para processamento distribuído (map, reduce, outras)
7. Conclusões (limitações e vantagens)
8. Referências

3. ARQUITETURA





https://online.visual-paradigm.com

4. INFRAESTRUTURA

Detalhes da configuração

Rótulo da versão: emr-5.33.1

Distribuição do Hadoop: Amazon 2.10.1

Aplicativos: Spark 2.4.7, Livy 0.7.0, Hive 2.3.7, JupyterEnterpriseGateway 2.1.0

URI do log: s3://aws-logs-009729866344-us-east-1/elasticmapreduce/

Visualização consistente do EMRFS: Desativado

ID personalizado de AMI: --

Application user interfaces

Serviço de histórico: Spark history server, YARN timeline server, Tez UI

Conexões: Not Enabled Habilitar conexão da web

Configuração do hardware

Tipo de instância m4.large — O tipo de instância selecionado adiciona um volume do EBS GP2 de 32 GiB padrão por instância. Saiba mais

Número de instâncias 3 (1 principal e 2 nós core)

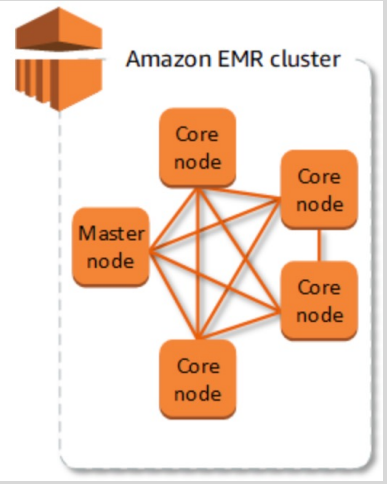
Escalabilidade do cluster Scale cluster nodes based on workload

Encerramento automático Habilitar encerramento automático Saiba mais

Encerrar o cluster quando ele estiver ocioso depois de 1 horas 0 minutos

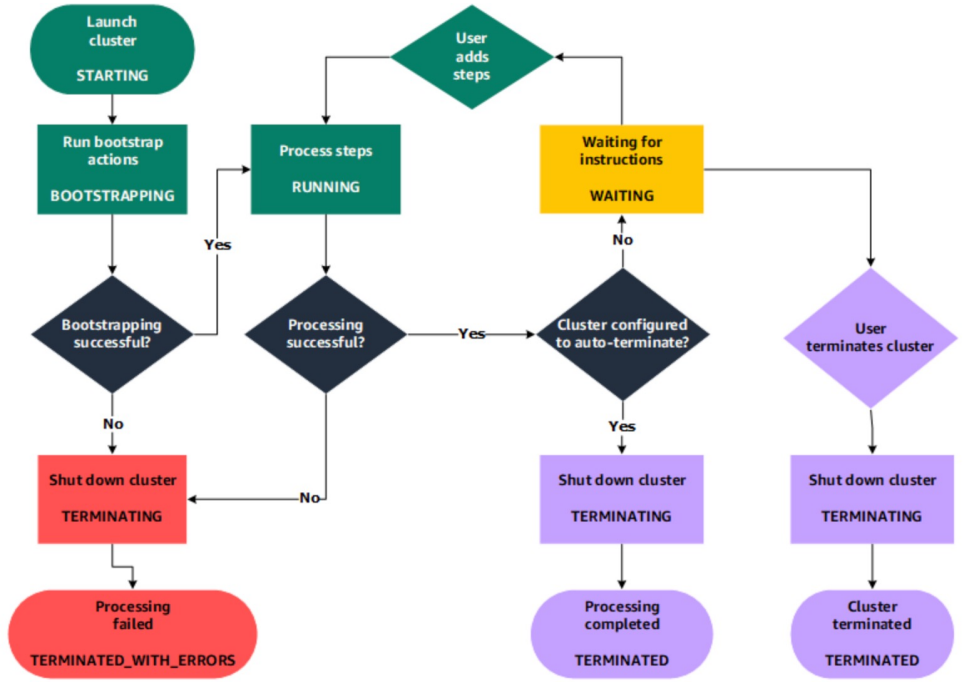
Dicas para criação do cluster

- Selecione as opções sugeridas pela AWS, são sempre mais rápidas e mais baratas
- Para testar seu trabalho, use um *cluster standalone*, que é um *cluster* com apenas um servidor. Depois de testado, vc pode testar seu trabalho em outras configurações.
- Abaixo, apresenta-se um cluster com 3 nós, um nó master(sempre) e 2 nós core. Veja mais detalhes aqui



Tipo e nome do nó	Tipo de instância	Contagem de instâncias
CORE Core Instance Group	**m4.large** 2 vCore, 8 GiB de memória, armazenamento apenas EBS Armazenamento de EBS: 32 GiB	2 Instâncias Redimensionar
MASTER Master Instance Group	**m4.large** 2 vCore, 8 GiB de memória, armazenamento apenas EBS Armazenamento de EBS: 32 GiB	1 Instâncias

CICLO DE VIDA DE UM CLUSTER



A menos que ocorra um erro, ou
que o cluster esteja configurado para encerramento automático AUTO TERMINATING,
o usuário tem sempre que ENCERRAR o cluster.

5. Análise do Processamento

Métricas	StandAlone Cluster	EMR multi node Cluster
ETL		
Modelo		
Resultados intermediários/ Armazenamento		

Pode ser uma comparação do processamento entre diferentes configurações de cluster, tais como: diferentes máquinas e diferentes números de nós.
Em ambientes gerenciados, como Sagemaker, não dá para mudar a configuração do cluster.
Por questões de custos, este tópico não é obrigatório.
Ademais, com as opções de *autoscaling o EMR* pode indicar quanto e quando aumentar os recursos de um cluster

8.Exemplos de notebooks

- Notebooks citados na especificação do trabalho
  - https://github.com/Mgosi/Big-Data-Analysis-using-MapReduce-in-Hadoop#readme
  - https://github.com/faiderfl/Big-Data--architecture-aws-spark#readme
- Notebooks do Sagemaker- Spark
  - https://github.com/aws/amazon-sagemaker-examples/tree/master/sagemaker-spark/pyspark_mnist