AWS EMR Best Practices and Resources

AWS Big Data Blog - HomePage for EMR Content

- https://aws.amazon.com/blogs/big-data/category/analytics/amazon-emr/?sc_ichannel=ha&sc_icampaign=acq_awsblogsb&sc_icontent=bigdata-resources

CloudFormation Templates

- Deploy EMR Cluster using CloudFormation Template
  - https://github.com/awslabs/aws-cloudformation-templates/blob/master/aws/services/EMR/EMRCLusterGangliaWithSparkOrS3backedHbase.json
- ETL Workflow using AWS MWAA (Airflow), Step Functions, Glue and EMR
  - https://aws.amazon.com/blogs/big-data/building-complex-workflows-with-amazon-mwaa-aws-step-functions-aws-glue-and-amazon-emr/
  - GitHub repo: https://github.com/aws-samples/amazon-mwaa-complex-workflow-using-step-functions
- ETL Workflow using AWS MWAA (Airflow) and EMR Notebooks
  - https://aws.amazon.com/blogs/big-data/orchestrating-analytics-jobs-on-amazon-emr-notebooks-using-amazon-mwaa/
- ETL Workflow with Airflow (Standalone) and Genie REST API in front of EMR Clusters
  - https://aws.amazon.com/blogs/big-data/orchestrate-big-data-workflows-with-apache-airflow-genie-and-amazon-emr-part-1/
  - https://aws.amazon.com/blogs/big-data/orchestrate-big-data-workflows-with-apache-airflow-genie-and-amazon-emr-part-2/
- ETL Workflow with Step Functions w/CloudFormation+AWS SAM
  - https://aws.amazon.com/blogs/big-data/automating-emr-workloads-using-aws-step-functions/
  - https://github.com/aws-samples/automation-ml-step-data-pipeline
- Building an EMR Self-Service Catalog
  - https://aws.amazon.com/blogs/big-data/build-a-self-service-environment-for-each-line-of-business-using-amazon-emr-and-aws-service-catalog/
- Monitor and Optimize EMR Workloads on AWS EMR using Prometheus and Grafana
  - https://aws.amazon.com/blogs/big-data/monitor-and-optimize-analytic-workloads-on-amazon-emr-with-prometheus-and-grafana/
- Monitor Spark Streaming Applications with CloudWatch
  - https://aws.amazon.com/blogs/big-data/monitor-spark-streaming-applications-on-amazon-emr/
- Accessing EMR Web interface using ALB Front-End
  - https://aws.amazon.com/blogs/big-data/access-web-interfaces-securely-on-amazon-emr-launched-in-a-private-subnet-using-an-application-load-balancer/

- Automate EMR DNS Registration with Route53 using CloudWatch Events and Lambda Functions
  - https://aws.amazon.com/blogs/big-data/dynamically-create-friendly-urls-for-your-amazon-emr-web-interfaces/

EMR Security Best Practices

- https://aws.amazon.com/blogs/big-data/best-practices-for-securing-amazon-emr/

EMR HBase Migration to S3

- https://aws.amazon.com/blogs/big-data/tips-for-migrating-to-apache-hbase-on-amazon-s3-from-hdfs/