Cloudera와 EMR 비교

1. AWS Migration Strategies

Preferred Strategies for Hadoop to EMR Migration

Lift and Shift

This strategy helps organizations achieve Hadoop to EMR migration faster to accelerate decommissioning of their on-premises data center. This enables organizations to eliminate cost-intensive hardware upgrades. The lift and shift strategy guides organizations to keep their existing Hadoop segregated and classified by utilizing AWS S3. Additionally, it helps them in decoupling resources, limiting code transformations to bare minimum. The simple lift and shift Hadoop to EMR migration approach moves the code as is to the cloud environment.

Replatform

The replatform strategy for Hadoop to EMR migration enables organizations to maximize their cloud migration advantages. This is basically done by utilizing the entire set of features provided by AWS EMR. With this strategy, organizations can fine tune their workloads and infrastructure for cost-effectiveness, scalability, and performance. Additionally, this strategy allows organizations to integrate their Hadoop ecosystem with cloud monitoring and security. Although replatform is similar to the lift and shift approach, it offers relatively lesser optimizations when it comes to cloud features and offerings.

Re-Architect

The strategy of re-architecting Hadoop on AWS EMR helps organizations to re-imagine their ecosystem of insights in the cloud. It helps them democratize their data to a larger customer pool while reducing the time-to-insight. This can be primarily attributed to the capabilities of streaming analytics, which provides organizations to self-service their requirements while building greater capabilities.

The re-architect strategy covers resolving all challenges of organizations, ranging from the analysis of business priorities to building a cloud-based data platform. The strategy basically involves changing the architecture with the help of cloud-native services to enhance performance, provision scalable solutions, and improve cost effectiveness of the infrastructure.

	AWS EMR	EC2의 CLOUDERA
Auto Scaling	EMR은 슬레이브 노드를 코어 노드와 작업 노드의 두 가지 하위 유형으로 분리합니다. 그 결과 작업 노드에 스팟 인스턴스를 사용하여 높은 확장 성과 낮은 비용이 발생합니다.	Cloudera 는 슬레이브 노드를 코어 및 작업 노드로 분류하지 않습니다. 따라서 노드 가 제거 / 분실 되면 HDFS 데이터 손실 위험이 증가합니다.
동적 오케스트레이션	매우 짧은 시간 내에 주문형 새 클러스터를 동적으로 오케스트레이션 할 수 있습니다. 이 클러스터는 작업이 성공적으로 완료된 후 종료 될 수 있습니다. 이렇게하면 활용도를 높일 수있어 비용을 크게 줄일 수 있습니다.	애플리케이션이 이미 ec2에서 실행중인 경우 불필요하게 리소스를 사용합니다. 비용을 절약하려면 데이터 처리를 위해 인스턴스를 시작 / 중지해야합니다.
Amazon S3에 대한 액세스	EMR에서 직접 또는 Hive 테이블을 통해 S3의 데이터에 액세스 할 수 있습니다. EMR은 AWS 독점 바이너리를 통해 S3의 데이터 작업을 위해 고도로 조정되었습니다. EMR은 Amazon Kinesis , Amazon Redshift 및 Amazon DynamoDB 와 같은 다른 Amazon 서비스와 원활하게 작동합니다 .	Cloudera는 Apache 라이브러리 (s3a)를 사용하여 S3의 데이터에 액세스하지만 EMR은 AWS 독점 코드를 사용하여 S3에 더 빠르게 액세스합니다.
고 가용성	EMR 서비스는 슬레이브 노드를 모니터링하고 비정상 노드를 새 노드로 교체합니다.	EMR과 달리 Cloudera는 슬레이브 노드를 코어 및 작업 노드로 분류하지 않습니다. 이는 노드가 제거 / 분실 된 경우 HDFS 데이터 손실 위험을 증가시킵니다.

2. cloudera vs aws emr 차이점

	AWS EMR	EC2의 CLOUDERA
Auto Scaling	EMR은 슬레이브 노드를 코어 노드와 작업 노드의 두 가지 하위 유형으로 분리합니다. 그 결과 작업 노드에 스팟 인스턴스를 사용하여 높은 확장 성과 낮은 비용이 발생합니다.	Cloudera 는 슬레이브 노드를 코어 및 작업 노드로 분류하지 않습니다. 따라서 노드 가 제거 / 분실 되면 HDFS 데이터 손실 위험이 증가합니다.
동적 오케스트레이션	매우 짧은 시간 내에 주문형 새 클러스터를 동적으로 오케스트레이션 할 수 있습니다. 이 클러스터는 작업이 성공적으로 완료된 후 종료 될 수 있습니다. 이렇게하면 활용도를 높일 수있어 비용을 크게 줄일 수 있습니다.	애플리케이션이 이미 ec2에서 실행중인 경우 불필요하게 리소스를 사용합니다. 비용을 절약하려면 데이터 처리를 위해 인스턴스를 시작 / 중지해야합니다.
Amazon S3에 대한 액세스	EMR에서 직접 또는 Hive 테이블을 통해 S3의 데이터에 액세스 할 수 있습니다. EMR은 AWS 독점 바이너리를 통해 S3의 데이터 작업을 위해 고도로 조정되었습니다. EMR은 Amazon Kinesis , Amazon Redshift 및 Amazon DynamoDB 와 같은 다른 Amazon 서비스와 원활하게 작동합니다 .	Cloudera는 Apache 라이브러리 (s3a)를 사용하여 S3의 데이터에 액세스하지만 EMR은 AWS 독점 코드를 사용하여 S3에 더 빠르게 액세스합니다.
고 가용성	EMR 서비스는 슬레이브 노드를 모니터링하고 비정상 노드를 새 노드로 교체합니다.	EMR과 달리 Cloudera는 슬레이브 노드를 코어 및 작업 노드로 분류하지 않습니다. 이는 노드가 제거 / 분실 된 경우 HDFS 데이터 손실 위험을 증가시킵니다.

3. Billing

비용 계산

Amazon EMR VS Cloudera 데이터 세트 처리를 테스트하기 위해 데이터 처리를위한 예제 (6 노드 Hadoop 클러스터 구성)를 살펴 보겠습니다 . 아래 계산은 미국 동부 지역에 대한 것이며 여기 에서 자세한 내용을 볼 수 있습니다 .

	AWS EMR	EC2의 CLOUDERA
1 년 동안 필요한 인스턴스	시간당 0.030 USD * 6 * 24 * 365	시간당 0.120 USD * 6 * 24 * 365
총 비용	1576.8 달러	$ 6307.2

우리는 일년 내내 빅 데이터 세트 처리를 실행해야하는 최악의 시나리오를 취했습니다. EMR의 경우 1 개의 마스터 노드와 5 개의 슬레이브 노드를 가질 수 있습니다. Cloudera의 경우 6 개의 노드를 실행하려면 6 개의 EC2 인스턴스가 필요합니다. 예약 된 인스턴스를 구입하여 비용을 절감 할 수 있지만 대부분의 시나리오에서는 이러한 클러스터가 더 짧은 기간 동안 필요 합니다. 따라서 위의 계산은 Cloudera를 사용하는 코어 EC2 클러스터에 비해 EMR이 매우 저렴하다는 것을 나타 냅니다.

결론

Amazon EMR VS Cloudera, 귀하의 선택은 특정 비즈니스 사례에 따라 다릅니다. 결과적으로 다음과 같은 경우 선택 사항이 있습니다.

- 배포 관리 및 업데이트에 시간을 투자하고 싶지 않다면 AWS EMR이 최상의 옵션이어야합니다.
- 데이터는 S3에 저장되고 데이터에 대해 가끔 작업을 실행하고 결과를 S3에 다시 덤프하려는 경우 Elastic Map / Reduce (EMR)를 사용하는 것이 합리적이어야합니다.