**A Comparative Analysis of Big Data Processing Using Hadoop and Spark on Cloud Platforms**

AWS EMR Cluster created

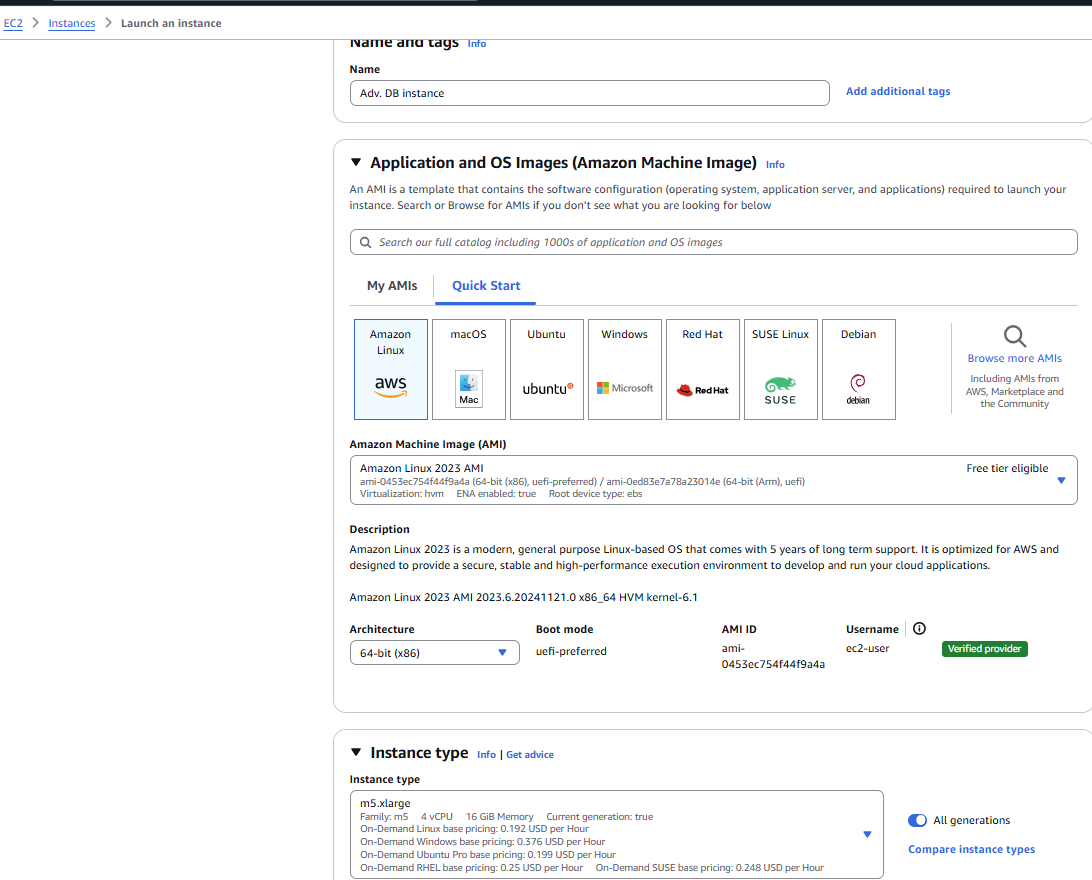
A computer screen with a computer screen and a computer screen

Description automatically generated

**A screenshot of a computer

Description automatically generated**

Data stored on S3 bucket



EC2 instance config. During creation

Databricks cluster created and processing of Task2 on it

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* It doesn't correspond directly to a specific amount of data, RAM, or CPU cycles. Instead, it reflects how much compute power (based on the type of virtual machine and workload) is being used on Databricks.

The **total cost** for running a job on Databricks can be calculated as: Total Cost=VM Cost+DBU

**VM Cost**: The cost of the specific VM instance you are using, based on its hourly rate.

* **DBU Cost**: Calculated as:
* DBU Cost= (DBU Count per VM×DBU Hourly Rate) × Time in Hours

**Cost Analysis for AWS**

**Amazon S3 Storage Costs:**

* **Storage:** Storing 1 TB (1,024 GB) in the S3 Standard storage class costs approximately $0.023 per GB per month.

**Calculation:** 1,024 GB \* $0.023/GB = $23.55 per month.

**Amazon EC2 Costs:**

**EC2 Instance Costs:** The primary cost component for EMR is the EC2 instances. Prices vary based on instance type and region.

* **Example:** An m5.xlarge instance (4 vCPUs, 16 GiB memory) in the US East (N. Virginia) region costs approximately $0.192 per hour.

**Cost Estimation Example:** Assuming you use a cluster of 10 m5.xlarge instances running for 10 hours:

* **EC2 Cost:** 10 instances \* $0.192/hour \* 10 hours = $19.20
* **EMR Cost:** EMR charges an additional fee per instance. For example, if the EMR fee is $0.027 per instance-hour:
  + 10 instances \* $0.027/hour \* 10 hours = $2.70
* **Total EMR Cluster Cost:** $19.20 (EC2) + $2.70 (EMR) = $21.90

**Total Estimated Cost:**

* **S3 Storage (monthly):** $23.55
* **EMR Cluster (for the job):** $21.90
* **Total:** $45.45

**Data for report included assumptions**

**Amazon S3 Storage Costs**

* **Storage:** Storing 1 TB (1,024 GB) in the S3 Standard storage class costs $0.023 per GB per month.  
  **Calculation:** 1,024 GB × $0.023/GB = **$23.55 per month.**
* **S3 Data Requests:**
  + **PUT, COPY, POST, LIST requests:** 1,000 requests cost $0.005. Assuming 10,000 PUT requests for 1 TB of data:  
    **10,000 × $0.005/1,000 = $0.05.**
  + **GET requests:** For 10,000 GET requests at $0.0004 per 1,000 requests:  
    **10,000 × $0.0004 = $0.004.**
  + **Total Request Costs:** $0.05 + $0.004 = **$0.054.**

**Amazon EC2 Costs**

* **Instance Type:** The default instance type for general-purpose workloads is **m5.xlarge** (4 vCPUs, 16 GiB memory).
* **Region:** Costs are based on the **US East (N. Virginia)** region.
* **EC2 On-Demand Price:** $0.192/hour per instance.
* **Cluster Configuration:**
  + 1 Master Node + 9 Core Nodes (10 instances total).
  + Estimated runtime: **10 hours**.
* **Cost Calculation for EC2 Instances:**
  + **EC2 Cost:** 10 instances × $0.192/hour × 10 hours = **$19.20.**

**Amazon EMR Costs**

* **EMR Service Fee:** $0.027 per instance-hour.
  + **EMR Fee Calculation:** 10 instances × $0.027/hour × 10 hours = **$2.70.**

**Data Transfer Costs**

* **Data Transfer Between S3 and EMR:** Free when data is transferred within the same AWS region.
* **Data Transfer Out to the Internet:** Not applicable for this use case.

**Total Estimated Cost for Processing the Job**

1. **S3 Storage Costs:**
   * Storage for 1 TB of data: **$23.55/month.**
   * Data requests: **$0.054.**

**Total S3 Cost:** $23.60.

1. **EMR Cluster Costs:**
   * EC2 instances: **$19.20.**
   * EMR service fee: **$2.70.**

**Total EMR Cost:** $21.90.

**Grand Total Cost:**

Total Cost=S3 Cost + EMR Cost=23.60+21.90=45.50 USD

**Assumptions Made:**

1. Data is stored and processed within the same AWS region to eliminate data transfer costs.
2. The cluster uses **m5.xlarge** instances (suitable for general-purpose workloads).
3. Processing time is estimated to be 10 hours based on the nature of the workload.
4. Default S3 Standard storage class and request pricing.
5. No additional features like encryption or versioning that incur additional costs.

**Stats for ppt in brief:**

1. **S3 Storage Cost:** Storing 1 TB of data costs **$23.55/month** in the S3 Standard storage class.
2. **S3 Request Cost:** 10,000 PUT and GET requests cost **$0.054**.
3. **EC2 Instance Cost:** Running 10 m5.xlarge instances for 10 hours costs **$19.20**.
4. **EMR Service Fee:** 10 m5.xlarge instances for 10 hours cost **$2.70**.
5. **Total Cost for 1 TB Processing:** Storing and processing the file costs **$45.50**.

**Price for Databricks**

A screenshot of a computer

Description automatically generated

Code to connect S3 bucket to your databricks notebook

To estimate the cost of processing a 1 TB file stored in Amazon S3 using Databricks on AWS, we'll consider the following components:

1. **Amazon S3 Storage Costs**
2. **Databricks Compute Costs**
3. **Data Transfer Costs**

**1. Amazon S3 Storage Costs**

* **Storage:** Storing 1 TB (1,024 GB) in the S3 Standard storage class costs $0.023 per GB per month.

**Calculation:** 1,024 GB × $0.023/GB = $23.55 per month.

* **Requests:**
  + **PUT, COPY, POST, LIST Requests:** $0.005 per 1,000 requests. Assuming 10,000 such requests:
    - **Calculation:** (10,000 / 1,000) × $0.005 = $0.05
  + **GET Requests:** $0.0004 per 1,000 requests. Assuming 10,000 such requests:
    - **Calculation:** (10,000 / 1,000) × $0.0004 = $0.004
  + **Total Request Cost:** $0.05 (PUT, etc.) + $0.004 (GET) = $0.054

**2. Databricks Compute Costs**

Databricks charges are based on **Databricks Units (DBUs)**, which represent the processing capability per hour. The cost depends on the instance type, workload, and pricing tier.

* **Instance Selection:** For this estimation, we'll use the **Standard Plan** with an **m5.xlarge** instance (4 vCPUs, 16 GiB memory).
* **DBU Consumption:** An m5.xlarge instance consumes 1 DBU per hour.
* **Cluster Configuration:** Assuming a cluster with 1 Driver node and 9 Worker nodes (10 m5.xlarge instances total).
* **Runtime:** Estimated at 10 hours.

**Calculations:**

* **Total DBUs Consumed:** 10 instances × 1 DBU/hour × 10 hours = 100 DBUs

**Compute Cost:** 100 DBUs × $0.40/DBU = $40.00

**3. Data Transfer Costs**

* **Between S3 and Databricks:** Data transfer between S3 and Databricks within the same AWS region is typically free.

**Total Estimated Cost:**

* **S3 Storage:** $23.55 (monthly)
* **S3 Requests:** $0.054
* **Databricks Compute:** $40.00

**Grand Total:** $23.55 (S3 Storage) + $0.054 (S3 Requests) + $40.00 (Databricks Compute) = **$63.604**

**Assumptions:**

* Data is stored and processed within the same AWS region to avoid inter-region data transfer costs.
* The cluster runs for exactly 10 hours; actual runtime may vary based on workload complexity.
* S3 request counts are estimated; actual numbers may differ.

**For PPT**

1. **S3 Storage Cost:** Storing 1 TB of data costs **$23.55/month** in the S3 Standard storage class.
2. **S3 Request Cost:** 10,000 PUT and GET requests cost **$0.054**.
3. **Databricks Compute Cost:** Running a cluster with 10 m5.xlarge instances for 10 hours costs **$40.00**.
4. **Total Cost for 1 TB Processing:** Storing and processing the file on Databricks costs **$63.60**.

A graph of a comparison between a number of data bricks and a platform

Description automatically generated

**Final Costs:**

* **AWS EMR:** $45.50
* **Databricks:** $63.60

**Reasons for Higher Costs on Databricks:**

1. **Databricks Compute Costs:** Databricks' per-DBU pricing structure increases overall costs compared to EMR’s EC2-based charges, especially with multiple instances.
2. **Platform Features:** Databricks offers enhanced tools like collaborative notebooks and machine learning integrations, which might justify the additional cost for certain use cases.
3. **Scalability and Optimization:** AWS EMR allows for greater flexibility in optimizing instance types and costs, reducing total runtime costs.
4. **Storage Costs:** Both platforms rely on S3, so storage-related expenses remain the same.

**Benefits of AWS EMR over Databricks**

1. **Cost Efficiency:**
   * EMR is generally more cost-effective for large-scale data processing tasks, as it directly uses EC2 instance pricing with minimal service fees.
2. **Flexibility in Instance Selection:**
   * EMR provides granular control over instance types, allowing users to optimize costs and performance based on specific workload requirements.
3. **Seamless Integration with AWS Ecosystem:**
   * Tight integration with other AWS services (e.g., S3, Lambda, Athena) enhances workflows and reduces data transfer complexities.
4. **Customizability:**
   * EMR allows users to configure and manage Spark clusters with high flexibility, including custom AMIs and software configurations.
5. **Pay-As-You-Go:**
   * EMR charges are strictly based on the runtime and resources used, enabling better cost predictability.

**Benefits of Databricks over AWS EMR**

1. **Advanced Collaboration Tools:**
   * Databricks provides built-in collaborative notebooks for data engineering, machine learning, and analytics, facilitating team collaboration.
2. **Optimized Spark Performance:**
   * Databricks’ proprietary runtime enhances Spark performance, offering faster execution and reduced runtime costs in some scenarios.
3. **Simplified Management:**
   * Databricks handles cluster setup and tuning automatically, reducing the operational overhead for data teams.
4. **Integrated ML and Analytics:**
   * With Databricks, users can seamlessly transition from data preparation to machine learning model development within the same platform.
5. **Cross-Cloud Availability:**
   * Databricks operates across AWS, Azure, and GCP, providing flexibility for multi-cloud strategies.

**When to Choose Each Platform:**

**Choose AWS EMR if:**

* Cost is a critical factor, and you’re comfortable with configuring and managing clusters.
* Your workloads are tightly coupled with AWS services.
* You need greater flexibility in customizing Spark environments.

**Choose Databricks if:**

* You need a collaborative, integrated environment for analytics and machine learning.
* Faster Spark performance and reduced management overhead are priorities.
* Your teams are focused on data science and require built-in ML tools and visualization.

**Summary:**

While **AWS EMR** offers cost savings and flexibility for large-scale processing, **Databricks** excels in usability, collaboration, and advanced features for data science workflows. The choice depends on the specific priorities of the project and organization.