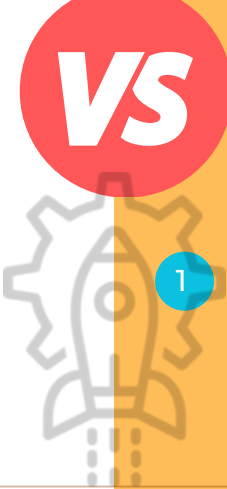


# HADOOP

VS

# AWS/AZURE

- 
- |   |  |   |   |
|---|--|---|---|
| 1 | Traditionally deployed on physical servers or on virtualized environments managed by the organization.   | 1 | Cloud-based platforms: Fully managed services, such as AWS EMR and Azure HDInsight, eliminate the need for physical infrastructure. |
| 2 | Requires significant upfront investment in hardware, networking, and storage infrastructure.             | 2 | Offer <b>on-demand scaling</b> with just a few clicks or automated policies (e.g., auto-scaling in EMR).                            |
| 3 | Maintenance and scaling depend entirely on the organization.   | 3 | No need to manage the underlying hardware; responsibility lies with the cloud provider.   |
| 4 | Organizations need to manage high availability and fault tolerance (e.g., configuring HDFS replication). | 4 | Built-in high availability and fault tolerance.   |

## 1. INFRASTRUCTURE AND DEPLOYMENT

# HADOOP

VS

# AWS/AZURE

1 CapEx-heavy: High initial costs for hardware and software licenses.

1 OpEx model: Pay-as-you-go pricing allows companies to pay only for what they use, making it cost-effective for short-term or variable workloads.



2 Recurring costs include maintenance, electricity, cooling, and personnel for managing the infrastructure.

2 Pricing is usage-based for compute, storage, and additional services.

3 Cost efficiency improves with long-term usage and heavy data processing needs.

3 Savings plans, such as spot instances (AWS) or reserved instances (Azure), can reduce costs for predictable workloads.

## 2. COST AND PRICING MODEL

# HADOOP

VS

# AWS/AZURE

- 1 Scaling is manual: Requires adding physical or virtual nodes to the cluster.
- 2 Upgrading capacity might require downtime, planning, and additional resources.
- 3 Performance is tied to on-premises limitations, such as power and cooling.

- 1 Seamless scalability: Cloud platforms support dynamic scaling based on workload requirements.
- 2 Horizontal scaling (adding nodes) can be done without downtime.
- 3 Virtually unlimited resources, subject to account quotas, making them ideal for large-scale operations.

## 3. SCALABILITY

# HADOOP

**VS**

# AWS/AZURE

- 1 Requires significant expertise in cluster setup, configuration, and management.
- 2 Uses a combination of tools like MapReduce, YARN, Hive, and Pig, which may have a steep learning curve.
- 3 Manual configuration for integrations with other tools (e.g., Apache Spark, Kafka).

- 1 Cloud services provide user-friendly interfaces for setup and management (e.g., AWS Management Console, Azure Portal).
- 2 Fully managed services handle installation, upgrades, and monitoring, reducing operational complexity.
- 3 Integrations with cloud-native tools (e.g., AWS Lambda, Azure Functions) are seamless and require minimal configuration.

## 4. EASE OF USE

# HADOOP

VS

# AWS/AZURE

- 1 Offers a rich ecosystem of tools (e.g., HDFS, YARN, Spark, Hive, HBase) for big data processing.
- 2 Integrations require manual effort, such as connecting to databases or visualization tools.
- 3 Often lacks native support for AI/ML frameworks and may need external tools for these tasks.

- 1 AWS EMR and Azure HDInsight include Hadoop-based frameworks but are also tightly integrated with cloud-native services.
  - AWS: S3, DynamoDB, SageMaker, Lambda.
  - Azure: Data Lake, Synapse Analytics, Machine Learning.
- 2 Pre-integrated AI/ML tools make deploying models faster and easier.
- 3 Native compatibility with visualization tools (e.g., AWS QuickSight, Power BI for Azure).

## 5. ECOSYSTEM AND INTEGRATION

# HADOOP

VS

# AWS/AZURE

- 1 Performance depends on hardware specifications and cluster configuration.
- 2 May require tuning and optimization of MapReduce jobs and HDFS for optimal performance.
- 3 Latency may increase with aging infrastructure or poorly optimized clusters.

- 1 **Optimized environments:** Cloud providers use high-performance hardware with options for custom tuning.
- 2 Performance benefits from proximity to **other services** (e.g., S3 for storage or RDS for databases in AWS).
- 3 Faster data processing due to advanced caching, tiered storage, and pre-optimized configurations.

## 6. PERFORMANCE

# HADOOP

VS

# AWS/AZURE

- 1 Security measures (e.g., Kerberos for authentication) must be implemented and managed manually.
- 2 Compliance with data regulations (e.g., GDPR, HIPAA) depends on the organization's infrastructure and policies.
- 3 Requires manual auditing and monitoring setups.

- 1 Cloud services **provide built-in security features**:
  - AWS: IAM, KMS, Shield.
  - Azure: Active Directory, Key Vault.
- 2 Automatic compliance with various global standards (e.g., SOC 2, ISO 27001).
- 3 Pre-configured **monitoring and alerting tools** for data breaches or anomalies.

## 7. SECURITY AND COMPLIANCE

# HADOOP

**VS**

# AWS/AZURE

- 1 Ideal for organizations with long-term, stable big data processing needs and skilled in-house teams.
- 2 Used for massive-scale batch processing and data lakes with predictable workloads.
- 3 Preferred by organizations that prioritize full control over their environment.

- 1 Best for organizations needing dynamic workloads, short-term projects, or burst processing.
- 2 Ideal for businesses already invested in cloud ecosystems or requiring global accessibility.
- 3 Great for integrating big data processing with AI/ML pipelines or serverless workflows.

## 8. USE CASES