



# **NetApp混合云数据解决方案 - 基于客户用例的 Spark 和 Hadoop**

NetApp artificial intelligence solutions

NetApp  
August 18, 2025

# 目录

|                                                   |    |
|---------------------------------------------------|----|
| NetApp混合云数据解决方案 - 基于客户用例的 Spark 和 Hadoop          | 1  |
| TR-4657: NetApp混合云数据解决方案 - 基于客户用例的 Spark 和 Hadoop | 1  |
| 为什么要进行 Hadoop 数据保护?                               | 1  |
| Hadoop 和 Spark 客户面临的数据保护挑战                        | 1  |
| 由NetApp提供支持的适用于大数据架构的数据结构                         | 2  |
| 经过验证的数据结构客户用例                                     | 3  |
| Hadoop 数据保护和NetApp                                | 5  |
| Hadoop 数据保护用例概述                                   | 5  |
| 用例 1: 备份 Hadoop 数据                                | 5  |
| 用例 2: 从云到本地的备份和灾难恢复                               | 5  |
| 用例 3: 在现有 Hadoop 数据上启用 DevTest                    | 6  |
| 用例 4: 数据保护和多云连接                                   | 6  |
| 用例 5: 加速分析工作负载                                    | 6  |
| 用例 1: 备份 Hadoop 数据                                | 6  |
| 要求和挑战                                             | 6  |
| 客户现有的备份解决方案x                                      | 6  |
| 备份解决方案                                            | 7  |
| 用例 2: 从云到本地的备份和灾难恢复                               | 10 |
| 场景                                                | 11 |
| 要求和挑战                                             | 11 |
| 解决方案                                              | 12 |
| 用例 3: 在现有 Hadoop 数据上启用 DevTest                    | 12 |
| 场景                                                | 12 |
| 要求和挑战                                             | 12 |
| 解决方案                                              | 13 |
| 用例 4: 数据保护和多云连接                                   | 13 |
| 场景                                                | 13 |
| 要求和挑战                                             | 14 |
| 解决方案                                              | 14 |
| 用例 5: 加速分析工作负载                                    | 14 |
| 场景                                                | 14 |
| 要求和挑战                                             | 15 |
| 解决方案                                              | 15 |
| 结束语                                               | 16 |
| 在哪里可以找到更多信息                                       | 16 |
| 致谢                                                | 16 |
| 版本历史                                              | 16 |

# NetApp混合云数据解决方案 - 基于客户用例的 Spark 和 Hadoop

## TR-4657: NetApp混合云数据解决方案 - 基于客户用例的 Spark 和 Hadoop

Karthikeyan Nagalingam 和 Sathish Thyagarajan, NetApp

本文档介绍了使用NetApp AFF和FAS存储系统、 NetApp Cloud Volumes ONTAP、NetApp互联存储以及适用于 Spark 和 Hadoop 的NetApp FlexClone技术的混合云数据解决方案。这些解决方案架构允许客户为其环境选择合适的数据保护解决方案。NetApp根据与客户及其业务用例的互动设计了这些解决方案。本文档提供以下详细信息：

- 为什么我们需要为 Spark 和 Hadoop 环境提供数据保护以及客户面临的挑战。
- 由NetApp愿景及其构建块和服务提供支持的数据结构。
- 如何使用这些构建块来构建灵活的数据保护工作流程。
- 根据实际客户使用案例分析几种架构的优缺点。每个用例提供以下组件：
  - 客户场景
  - 要求和挑战
  - 解决方案
  - 解决方案总结

### 为什么要进行 Hadoop 数据保护？

在 Hadoop 和 Spark 环境中，必须解决以下问题：

- **\*软件或人为故障。\***在执行 Hadoop 数据操作时，软件更新中的人为错误可能会导致错误行为，从而导致工作出现意外结果。在这种情况下，我们需要保护数据以避免失败或不合理的结果。例如，由于交通信号分析应用程序的软件更新执行不力，导致新功能无法正确分析纯文本形式的交通信号数据。该软件仍然分析JSON和其他非文本文件格式，导致实时交通管制分析系统产生缺少数据点的预测结果。这种情况可能会导致错误输出，从而引发交通信号事故。数据保护可以通过提供快速回滚到以前工作应用程序版本的功能来解决此问题。
- **\*尺寸和规模。\***由于数据源数量和数据量的不断增加，分析数据的大小也日益增长。社交媒体、移动应用、数据分析和云计算平台是当前大数据市场的主要数据来源，这些数据增长非常迅速，因此需要对数据进行保护，以确保数据操作的准确性。
- **Hadoop 的原生数据保护。** Hadoop 有一个原生命令来保护数据，但是该命令在备份期间不提供数据的一致性。它仅支持目录级备份。Hadoop 创建的快照是只读的，不能直接用于重复使用备份数据。

### Hadoop 和 Spark 客户面临的数据保护挑战

Hadoop 和 Spark 客户面临的一个共同挑战是减少备份时间并提高备份可靠性，同时又不会在数据保护期间对生产集群的性能产生负面影响。

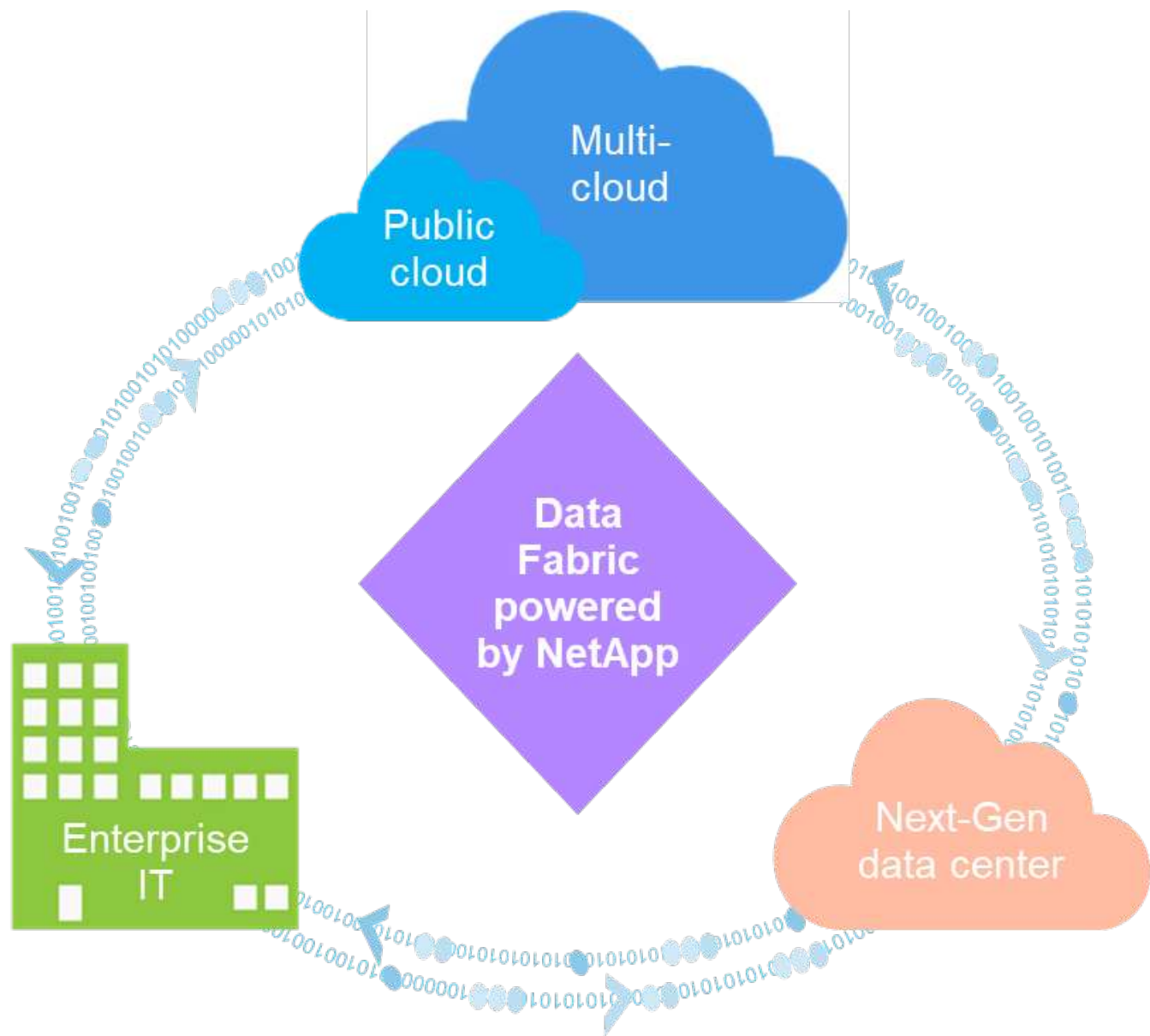
客户还需要最大限度地减少恢复点目标 (RPO) 和恢复时间目标 (RTO) 停机时间，并控制其内部部署和基于云的灾难恢复站点，以实现最佳业务连续性。这种控制通常来自于企业级管理工具。

Hadoop 和 Spark 环境非常复杂，因为不仅数据量巨大且不断增长，而且数据到达的速度也在加快。这种情况使得从源数据快速创建高效、最新的 DevTest 和 QA 环境变得困难。NetApp 认识到这些挑战并提供了本文中介绍的解决方案。

## 由NetApp提供支持的适用于大数据架构的数据结构

NetApp提供支持的数据结构简化并集成了跨云和本地环境的数据管理，从而加速数字化转型。

NetApp提供支持的数据结构为数据可见性和洞察、数据访问和控制以及数据保护和安全提供了一致且集成的数据管理服务 and 应用程序（构建块），如下图所示。



## 经过验证的数据结构客户用例

NetApp支持的数据结构为客户提供了以下九种经过验证的用例：

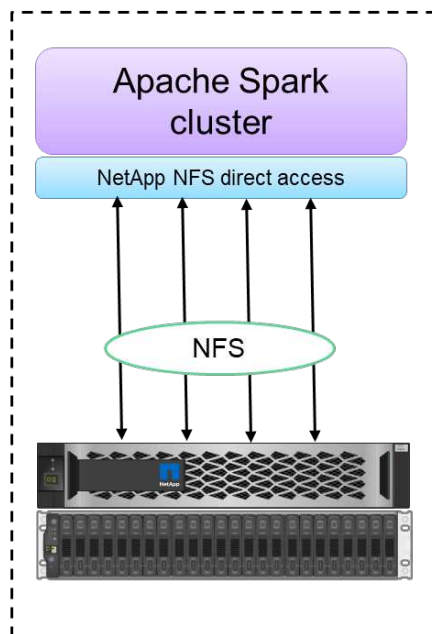
- 加速分析工作负载
- 加速 DevOps 转型
- 构建云托管基础设施
- 集成云数据服务
- 保护和保障数据安全
- 优化非结构化数据
- 提高数据中心效率
- 提供数据洞察和控制
- 简化和自动化

本文档涵盖九个用例中的两个（及其解决方案）：

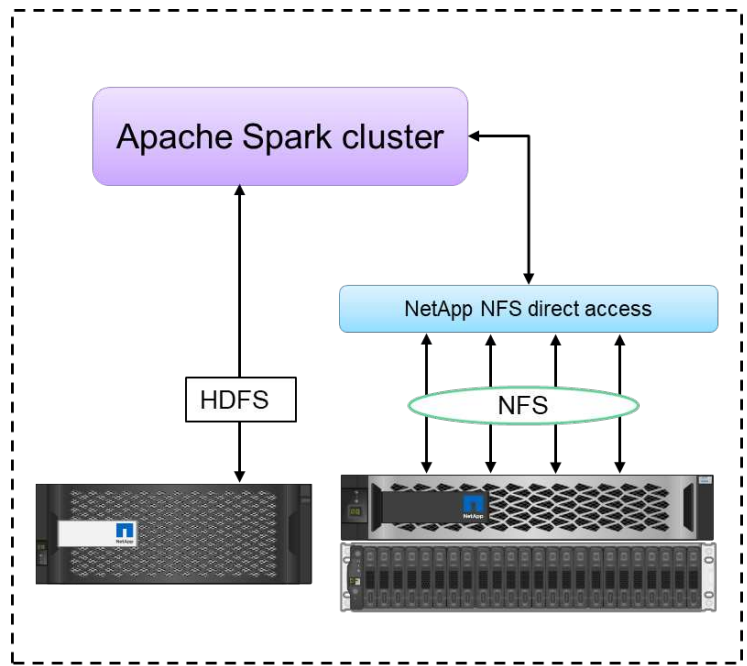
- 加速分析工作负载
- 保护和保障数据安全

### NetApp NFS 直接访问

NetApp NFS 允许客户在其现有或新的 NFSv3 或 NFSv4 数据上运行大数据分析作业，而无需移动或复制数据。它可以防止数据的多次复制，并且无需将数据与源同步。例如，在金融领域，数据从一个地方移动到另一个地方必须满足法律义务，这不是一件容易的事。在这种情况下，NetApp NFS 直接访问会从原始位置分析财务数据。另一个主要优势是，使用NetApp NFS 直接访问可以通过使用本机 Hadoop 命令简化 Hadoop 数据的保护，并利用 NetApp 丰富的数据管理产品组合实现数据保护工作流程。



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

NetApp NFS 直接访问为 Hadoop/Spark 集群提供了两种部署选项：

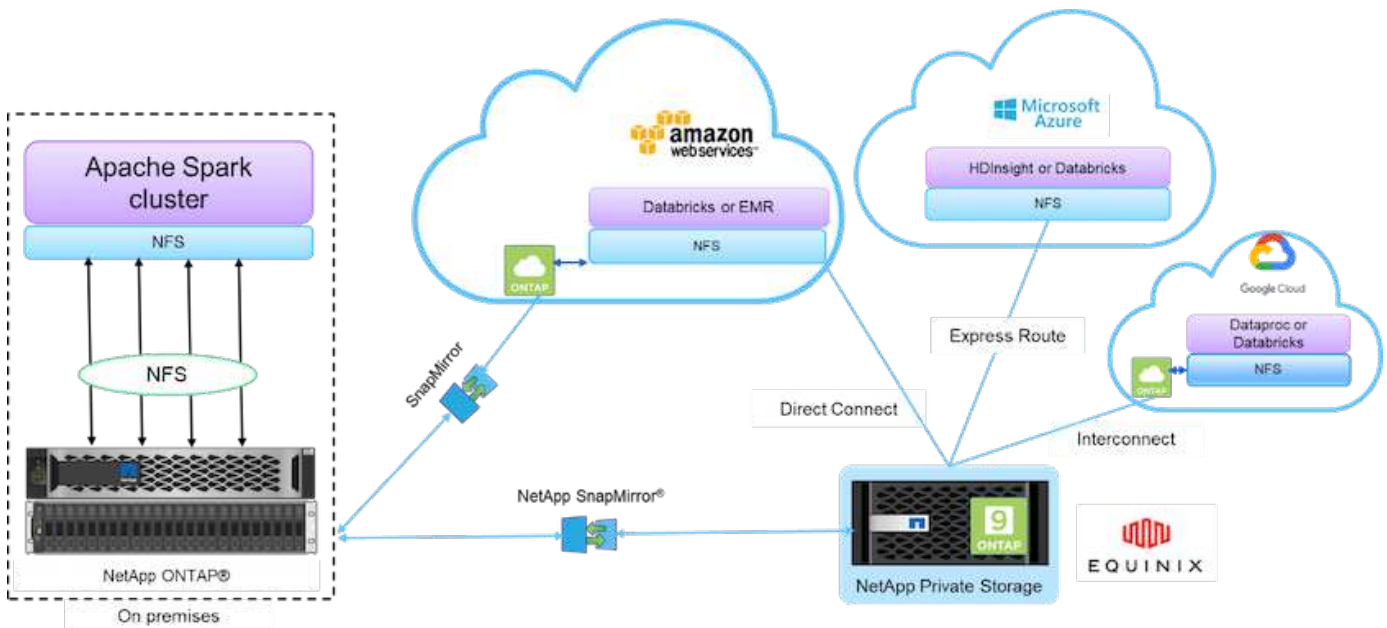
- 默认情况下，Hadoop/Spark 集群使用 Hadoop 分布式文件系统（HDFS）作为数据存储和默认文件系统。NetApp NFS 直接访问可以用 NFS 存储替换默认的 HDFS 作为默认文件系统，从而实现对 NFS 数据的直接分析操作。
- 在另一个部署选项中，NetApp NFS 直接访问支持在单个 Hadoop/Spark 集群中将 NFS 与 HDFS 一起配置为附加存储。在这种情况下，客户可以通过 NFS 导出共享数据，并从同一个集群访问数据以及 HDFS 数据。

使用 NetApp NFS 直接访问的主要优势包括：

- 从当前位置分析数据，从而避免将分析数据移动到 Hadoop 基础架构（如 HDFS）这一耗时耗能的任務。
- 将副本数量从三个减少到一个。
- 使用户能够分离计算和存储以独立扩展它们。
- 利用 ONTAP 丰富的数据管理功能提供企业数据保护。
- 已通过 Hortonworks 数据平台认证。
- 支持混合数据分析部署。
- 利用动态多线程功能减少备份时间。

## 大数据的构建模块

NetApp 提供支持的数据结构集成了数据管理服务 and 应用程序（构建块），用于数据访问、控制、保护和 safety，如下图所示。



上图中的构建块包括：

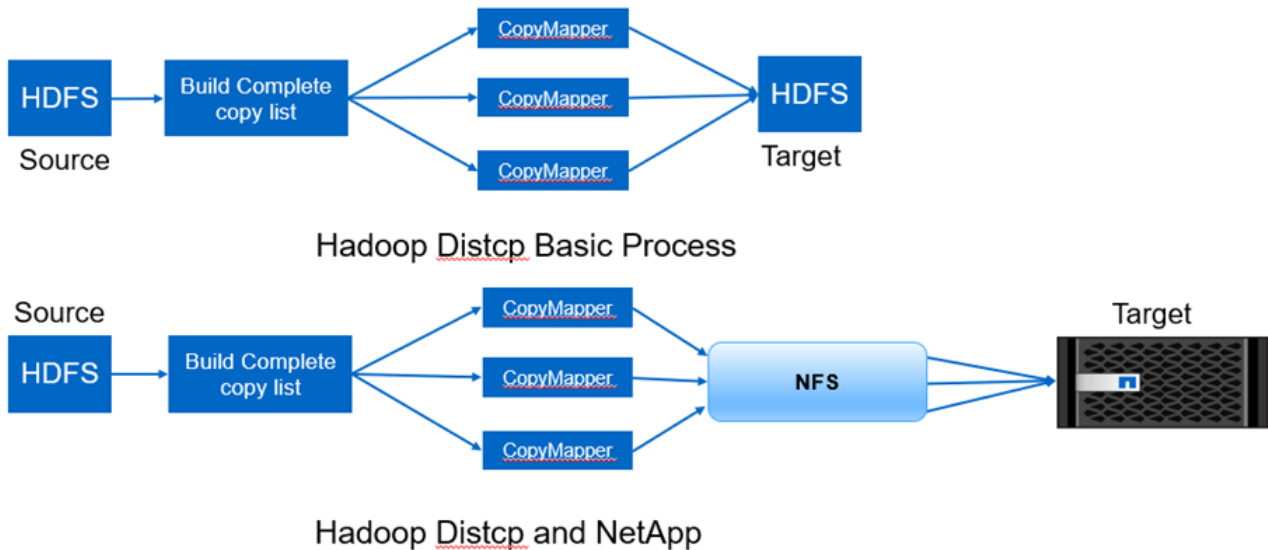
- \* NetApp NFS 直接访问。\* 为最新的 Hadoop 和 Spark 集群提供对 NetApp NFS 卷的直接访问，无需额外的软件或驱动程序要求。
- \* NetApp Cloud Volumes ONTAP 和 Google Cloud NetApp Volumes。\* 基于在 Amazon Web Services (AWS) 或 Microsoft Azure 云服务中的 Azure NetApp Files (ANF) 中运行的 ONTAP 的软件定义连接存储。

- \* NetApp SnapMirror技术\*。在本地和ONTAP Cloud 或 NPS 实例之间提供数据保护功能。
- \*云服务提供商\*。\*这些提供商包括 AWS、Microsoft Azure、Google Cloud 和 IBM Cloud。
- \*平台即服务 (PaaS)\*。\*基于云的分析服务，例如 AWS 中的 Amazon Elastic MapReduce (EMR) 和 Databricks 以及 Microsoft Azure HDInsight 和 Azure Databricks。

## Hadoop 数据保护和NetApp

Hadoop DistCp 是一个用于大型集群间和集群内复制的本机工具。下图所示的 Hadoop DistCp 基本流程是一个典型的备份工作流，使用 MapReduce 等 Hadoop 原生工具将 Hadoop 数据从 HDFS 源复制到相应的目标。

NetApp NFS 直接访问使客户能够将 NFS 设置为 Hadoop DistCp 工具的目标位置，以便通过 MapReduce 将数据从 HDFS 源复制到 NFS 共享中。NetApp NFS 直接访问充当 DistCp 工具的目标位置。



## Hadoop 数据保护用例概述

本节对数据保护用例进行了高层描述，这是本文的重点。其余部分为每个用例提供了更多详细信息，例如客户问题（场景）、要求和挑战以及解决方案。

### 用例 1：备份 Hadoop 数据

对于此用例，NetApp NFS 卷帮助一家大型金融机构将长备份窗口时间从 24 多个小时缩短至几个小时。

### 用例 2：从云到本地的备份和灾难恢复

通过使用NetApp提供支持的作为构建模块，一家大型广播公司能够根据不同的数据传输模式（例如按需、即时或基于 Hadoop/Spark 集群负载）满足将云数据备份到其内部数据中心的需求。

### 用例 3：在现有 Hadoop 数据上启用 DevTest

NetApp解决方案帮助一家在线音乐分销商在不同分支机构快速构建多个节省空间的 Hadoop 集群，以创建报告并通过使用计划策略运行日常 DevTest 任务。

### 用例 4：数据保护和多云连接

一家大型服务提供商使用由NetApp提供支持的数据结构从不同的云实例为其客户提供多云分析。

### 用例 5：加速分析工作负载

一家最大的金融服务和投资银行使用NetApp网络附加存储解决方案来减少 I/O 等待时间并加速其定量金融分析平台。

## 用例 1：备份 Hadoop 数据

在这种情况下，客户拥有一个大型的内部部署 Hadoop 存储库，并希望将其备份以用于灾难恢复目的。然而，客户当前的备份解决方案成本高昂，并且备份窗口长达 24 小时以上。

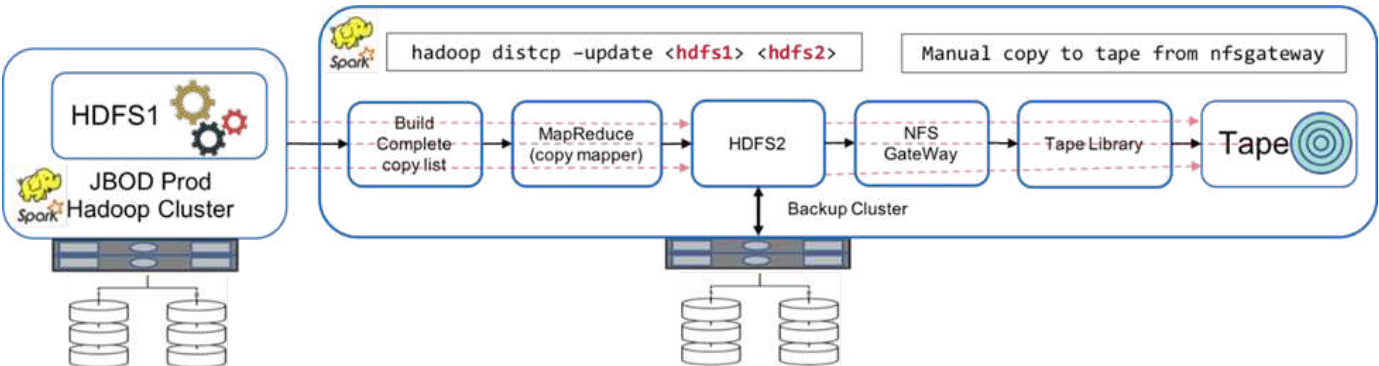
### 要求和挑战

此用例的主要要求和挑战包括：

- 软件向后兼容性：
  - 所提出的替代备份解决方案应与生产 Hadoop 集群中当前运行的软件版本兼容。
- 为了满足承诺的 SLA，建议的替代解决方案应该实现非常低的 RPO 和 RTO。
- NetApp备份解决方案创建的备份可用于在数据中心本地构建的 Hadoop 集群以及在远程站点的灾难恢复位置运行的 Hadoop 集群。
- 所提出的解决方案必须具有成本效益。
- 所提出的解决方案必须减少备份期间对当前正在运行的生产分析作业的性能影响。

### 客户现有的备份解决方案x

下图是原有的Hadoop原生备份方案。



生产数据通过中间备份集群保护到磁带上：

- 通过运行以下命令将 HDFS1 数据复制到 HDFS2 `hadoop distcp -update <hdfs1> <hdfs2>` 命令。
- 备份集群充当NFS网关，通过Linux手动将数据复制到磁带 `cp` 通过磁带库命令。

Hadoop原生备份方案的优势包括：

- 该解决方案基于 Hadoop 原生命令，使用户无需学习新的程序。
- 该解决方案利用行业标准架构和硬件。

原有Hadoop原生备份方案的缺点包括：

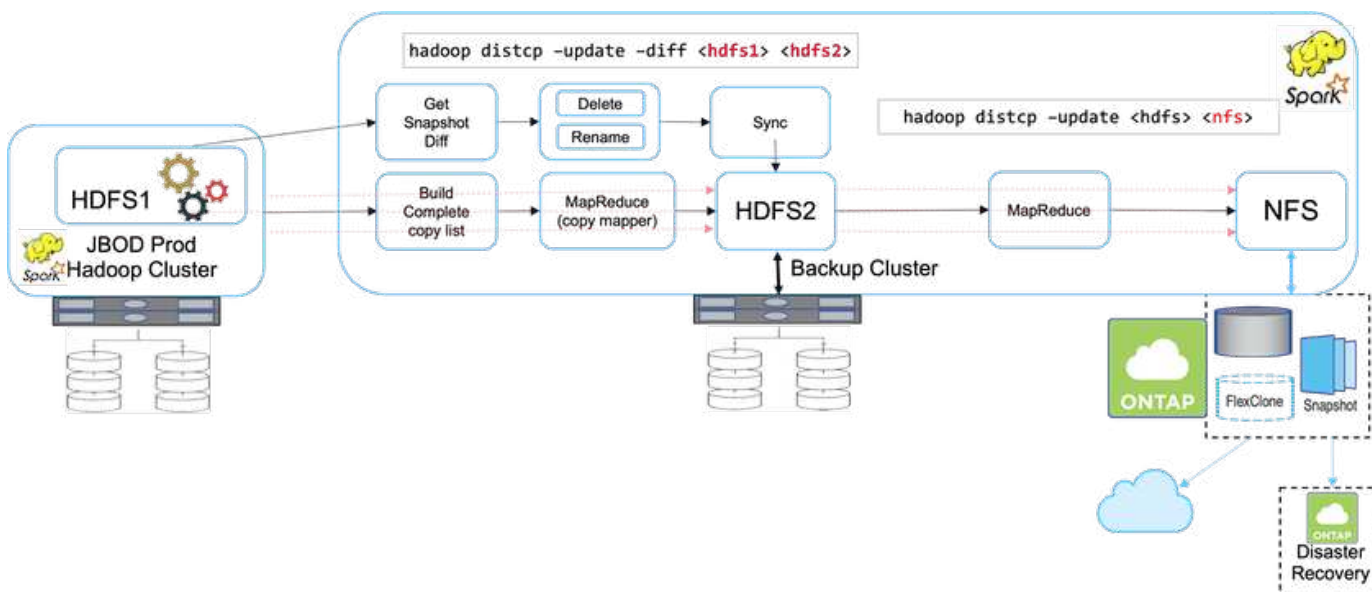
- 备份窗口时间过长超过24小时，导致生产数据容易受到攻击。
- 备份期间集群性能明显下降。
- 复制到磁带是一个手动过程。
- 就所需硬件和手动流程所需的人力而言，备份解决方案的成本很高。

## 备份解决方案

基于这些挑战和要求，并考虑到现有的备份系统，提出了三种可能的备份解决方案。以下小节分别描述这三种不同的备份解决方案，标记为解决方案 A 到解决方案 C。

### 解决方案 A

在解决方案 A 中，备份 Hadoop 集群将二级备份发送到NetApp NFS 存储系统，从而无需磁带，如下图所示。



解决方案A的详细任务包括：

- 生产 Hadoop 集群在需要保护的 HDFS 中拥有客户的分析数据。
- 带有 HDFS 的备份 Hadoop 集群充当数据的中间位置。只需一组磁盘（JBOD）即可为生产和备份 Hadoop 集群中的 HDFS 提供存储。

- 通过运行 `Hadoop distcp -update -diff <hdfs1> <hdfs2>` 命令。



Hadoop快照用于保护从生产到备份Hadoop集群的数据。

- NetApp ONTAP存储控制器提供 NFS 导出卷，该卷配置给备份 Hadoop 集群。
- 通过运行 `Hadoop distcp` 命令利用 MapReduce 和多个映射器，分析数据从备份 Hadoop 集群到 NFS 受到保护。

将数据存储在NetApp存储系统上的 NFS 中后，根据需要使用NetApp Snapshot、 SnapRestore和FlexClone 技术备份、恢复和复制 Hadoop 数据。



通过使用SnapMirror技术，Hadoop 数据可以受到保护，保存到云端以及灾难恢复位置。

解决方案A的优点包括：

- Hadoop 生产数据受到备份集群的保护。
- HDFS 数据通过 NFS 进行保护，从而实现对云和灾难恢复位置的保护。
- 通过将备份操作卸载到备份集群来提高性能。
- 消除手动磁带操作
- 允许通过NetApp工具实现企业管理功能。
- 只需对现有环境进行最少的改变。
- 是一种经济有效的解决方案。

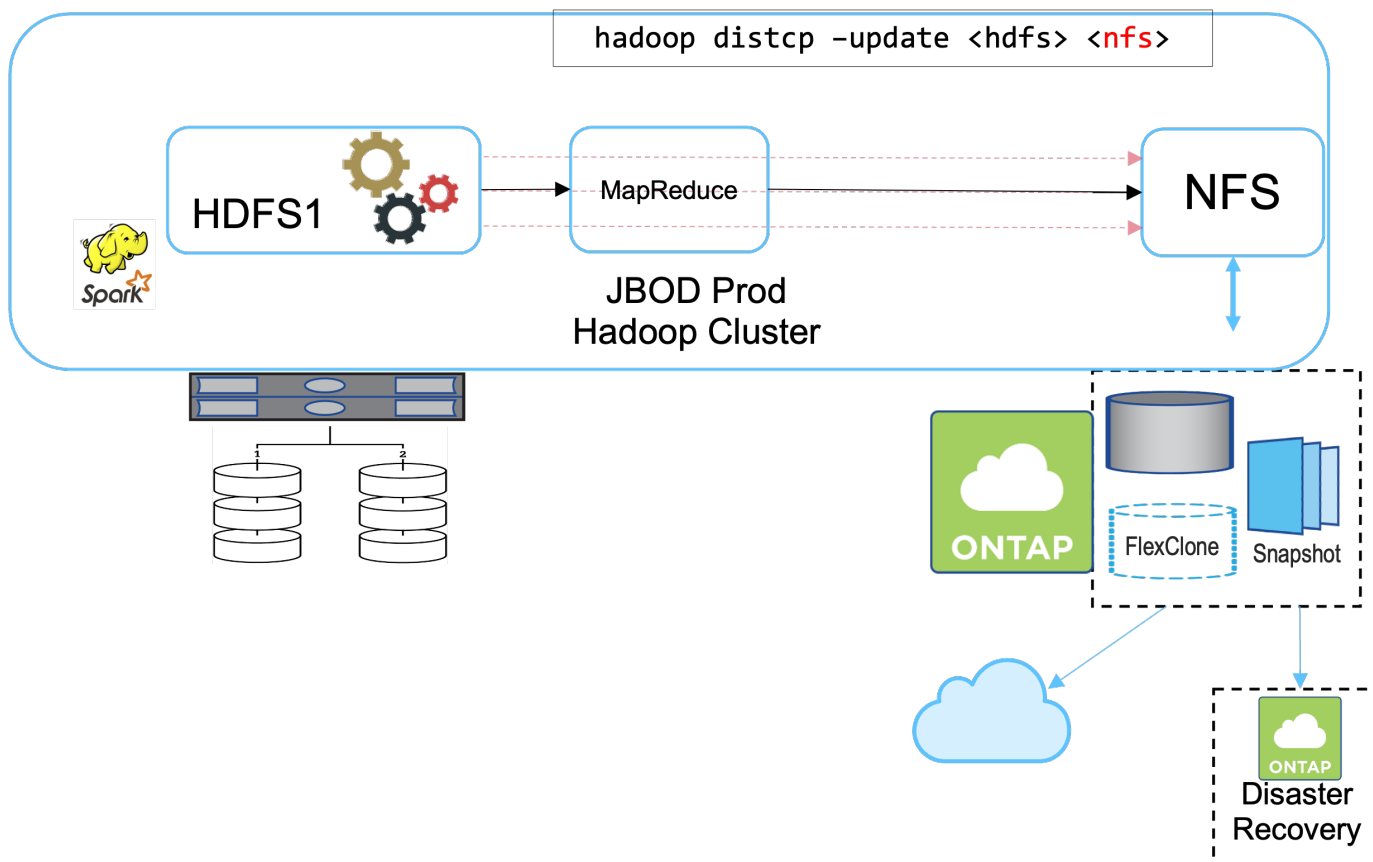
该解决方案的缺点是它需要备份集群和额外的映射器来提高性能。

客户最近部署了解决方案 A，因为它简单、成本低且整体性能好。

在此解决方案中，可以使用ONTAP的 SAN 磁盘代替 JBOD。此选项将备份集群存储负载转移至ONTAP；但缺点是需要 SAN 结构交换机。

## 解决方案 B

解决方案B将NFS卷添加到生产Hadoop集群，从而无需备份Hadoop集群，如下图所示。



解决方案B的详细任务包括：

- NetApp ONTAP存储控制器将 NFS 导出配置到生产 Hadoop 集群。

Hadoop 原生 `hadoop distcp` 命令将 Hadoop 数据从生产集群 HDFS 保护到 NFS。

- 将数据存储到NetApp存储系统的NFS中后，根据需要使Snapshot、 SnapRestore、 FlexClone技术对Hadoop数据进行备份、恢复、复制。

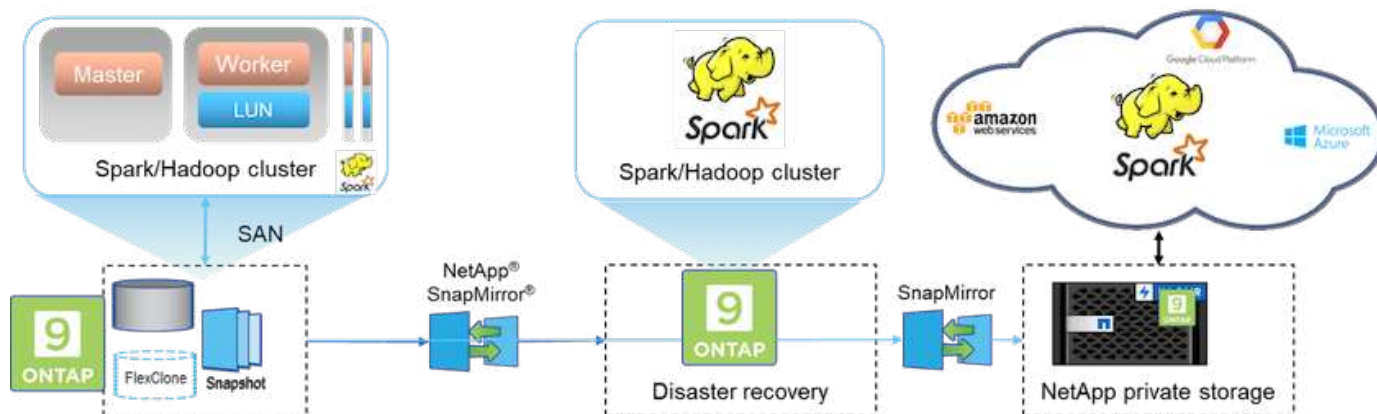
解决方案B的优点包括：

- 生产集群针对备份解决方案进行了轻微修改，简化了实施并降低了额外的基础设施成本。
- 备份操作不需要备份集群。
- HDFS 生产数据在转换为 NFS 数据时受到保护。
- 该解决方案允许通过NetApp工具实现企业管理功能。

该解决方案的缺点是它是在生产集群中实现的，这会在生产集群中增加额外的管理员任务。

### 解决方案 C

在解决方案 C 中， NetApp SAN 卷直接配置到 Hadoop 生产集群用于 HDFS 存储，如下图所示。



解决方案C的详细步骤包括：

- NetApp ONTAP SAN 存储在生产 Hadoop 集群中配置用于 HDFS 数据存储。
- NetApp Snapshot 和 SnapMirror 技术用于备份生产 Hadoop 集群的 HDFS 数据。
- 由于备份位于存储层，因此 Snapshot 复制备份过程中不会对 Hadoop/Spark 集群的生产性能产生影响。



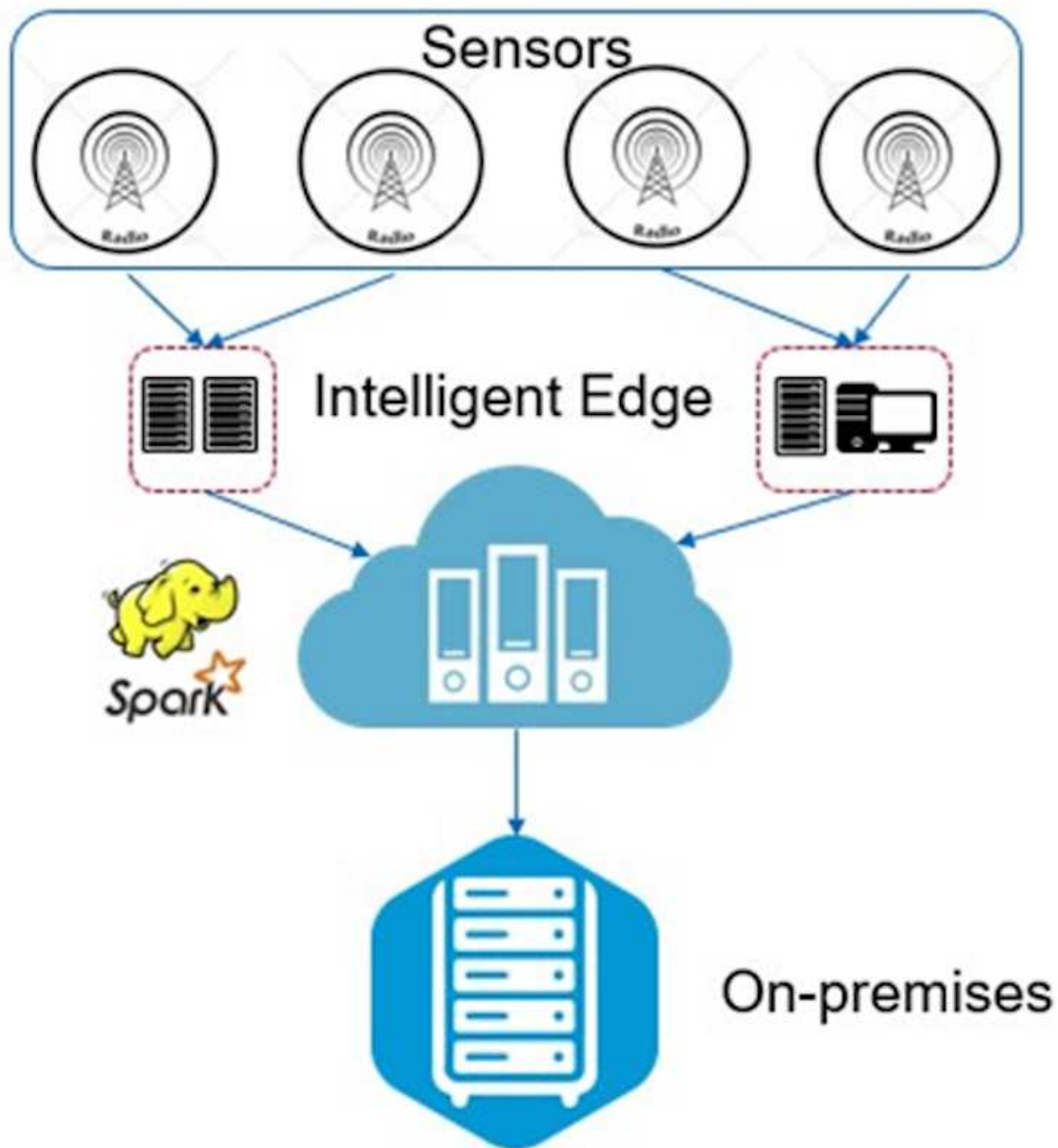
快照技术可提供在几秒钟内完成的备份，无论数据大小如何。

解决方案C的优点包括：

- 可以使用快照技术创建节省空间的备份。
- 允许通过 NetApp 工具实现企业管理功能。

## 用例 2：从云到本地的备份和灾难恢复

此用例基于一个广播客户，该客户需要将基于云的分析数据备份到其内部数据中心，如下图所示。



## 场景

在这种情况下，物联网传感器数据被引入云中，并使用 AWS 内的开源 Apache Spark 集群进行分析。要求是将处理后的数据从云端备份到本地。

## 要求和挑战

此用例的主要要求和挑战包括：

- 启用数据保护不会对云中的生产 Spark/Hadoop 集群造成任何性能影响。
- 需要以高效、安全的方式将云传感器数据移动并保护到本地。
- 在不同条件下灵活地将数据从云端传输到本地，例如按需、即时以及低集群负载时间期间。

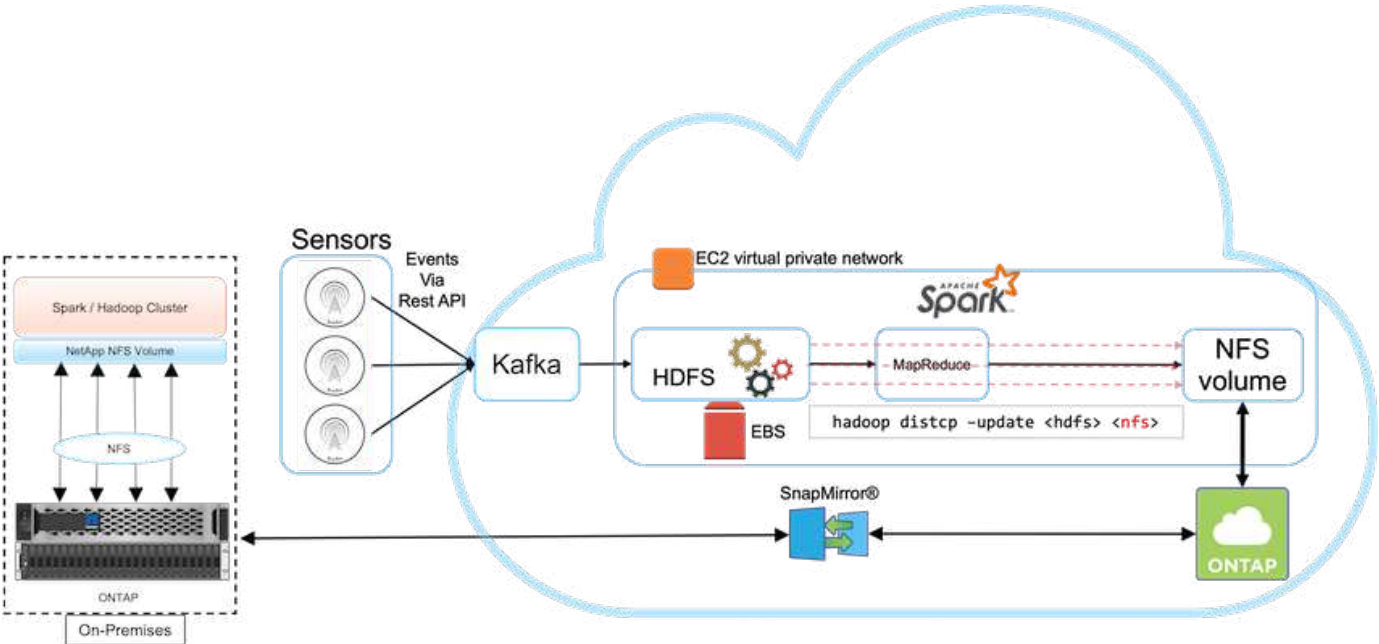
## 解决方案

客户使用 AWS Elastic Block Store (EBS) 作为其 Spark 集群 HDFS 存储，以通过 Kafka 接收和提取来自远程传感器的数据。因此，HDFS 存储充当备份数据的来源。

为了满足这些要求，NetApp ONTAP Cloud 部署在 AWS 中，并创建了一个 NFS 共享作为 Spark/Hadoop 集群的备份目标。

创建 NFS 共享后，将数据从 HDFS EBS 存储复制到 ONTAP NFS 共享。当数据驻留在 ONTAP Cloud 中的 NFS 中时，可以根据需要使用 SnapMirror 技术以安全高效的方式将数据从云端镜像到本地存储中。

此图显示了从云到本地解决方案的备份和灾难恢复。



## 用例 3：在现有 Hadoop 数据上启用 DevTest

在这种用例中，客户的要求是基于现有的 Hadoop 集群快速高效地构建新的 Hadoop/Spark 集群，该集群包含大量用于同一数据中心和远程位置的开发生态和报告目的的分析数据。

### 场景

在这种情况下，从本地以及灾难恢复位置的大型 Hadoop 数据湖实现构建多个 Spark/Hadoop 集群。

### 要求和挑战

此用例的主要要求和挑战包括：

- 为 DevTest、QA 或任何其他需要访问相同生产数据的目的创建多个 Hadoop 集群。这里的挑战是以非常节省空间的方式瞬间多次克隆一个非常大的 Hadoop 集群。
- 将 Hadoop 数据同步到 DevTest 和报告团队以提高运营效率。

- 在生产和新集群中使用相同的凭据分发 Hadoop 数据。
- 使用调度策略高效地创建QA集群，而不影响生产集群。

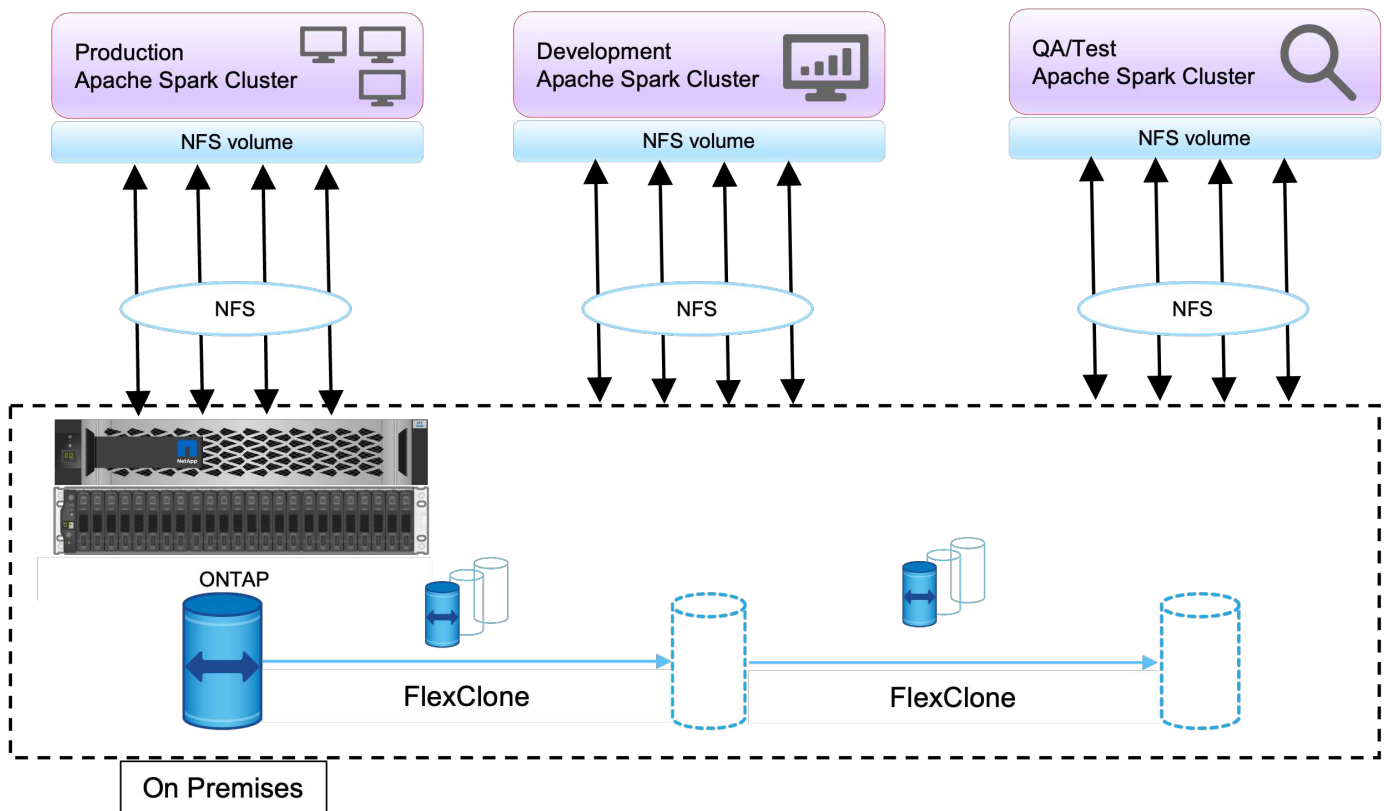
## 解决方案

FlexClone技术用于满足刚才描述的要求。FlexClone技术是 Snapshot 副本的读/写副本。它从父 Snapshot 副本数据中读取数据，并且仅为新/修改的块消耗额外的空间。它速度快并且节省空间。

首先，使用NetApp一致性组创建现有集群的 Snapshot 副本。

NetApp系统管理器或存储管理提示中的快照副本。一致性组Snapshot副本是应用程序一致性组Snapshot副本，FlexClone卷是基于一致性组Snapshot副本创建的。值得一提的是，FlexClone卷继承了父卷的 NFS 导出策略。创建 Snapshot 副本后，必须安装一个新的 Hadoop 集群以用于 DevTest 和报告目的，如下图所示。来自新 Hadoop 集群的克隆 NFS 卷访问 NFS 数据。

此图显示了 DevTest 的 Hadoop 集群。



## 用例 4：数据保护和多云连接

此用例与负责为客户的大数据分析数据提供多云连接的云服务合作伙伴相关。

### 场景

在这种情况下，AWS 从不同来源接收的物联网数据存储于 NPS 的中心位置。NPS 存储连接到位于 AWS 和 Azure 的 Spark/Hadoop 集群，使在多个云中运行的大数据分析应用程序能够访问相同的数据。

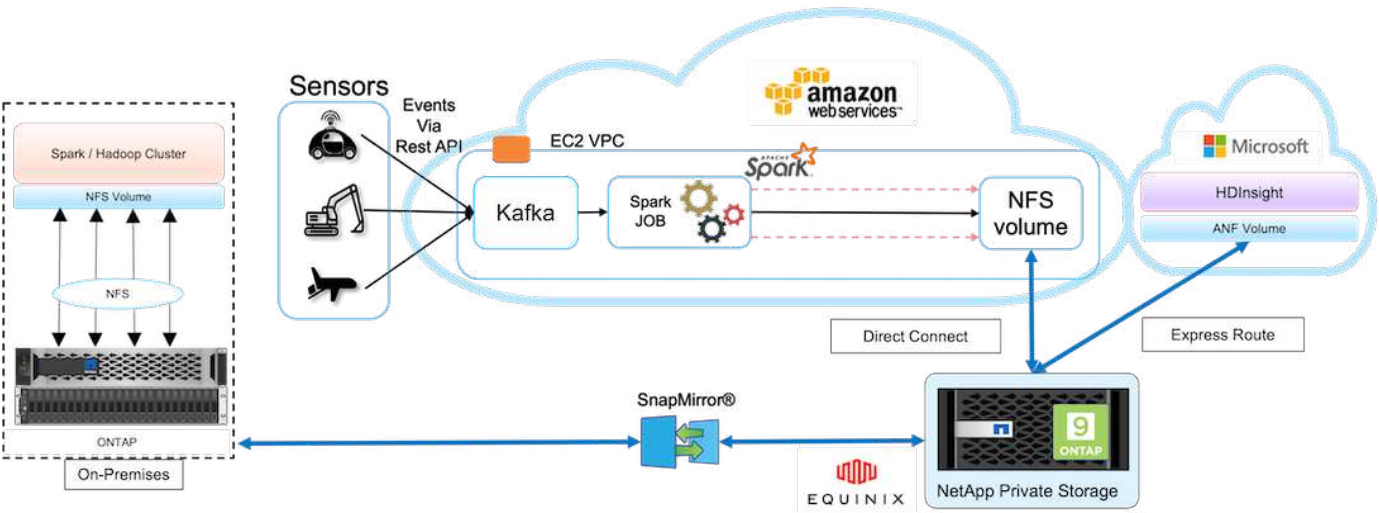
# 要求和挑战

此用例的主要要求和挑战包括：

- 客户希望使用多个云对相同数据运行分析作业。
- 必须通过不同的传感器和集线器从不同来源（例如本地和云端）接收数据。
- 该解决方案必须高效且具有成本效益。
- 主要挑战是构建一个经济高效的解决方案，在本地和不同云之间提供混合分析服务。

# 解决方案

此图说明了数据保护和多云连接解决方案。



如上图所示，来自传感器的数据通过 Kafka 流式传输并输入到 AWS Spark 集群中。数据存储在 NPS 中的 NFS 共享中，NPS 位于 Equinix 数据中心内的云提供商之外。由于 NetApp NPS 分别通过 Direct Connect 和 Express Route 连接连接到 Amazon AWS 和 Microsoft Azure，因此客户可以从 Amazon 和 AWS 分析集群访问 NFS 数据。这种方法解决了跨多个超大规模器的云分析问题。

因此，由于内部部署和 NPS 存储都运行 ONTAP 软件， SnapMirror 可以将 NPS 数据镜像到内部部署集群中，从而提供跨内部部署和多个云的混合云分析。

为了获得最佳性能， NetApp 通常建议使用多个网络接口和直接连接/快速路由来访问云实例的数据。

# 用例 5：加速分析工作负载

在此场景中，一家大型金融服务和投资银行的分析平台使用 NetApp NFS 存储解决方案进行了现代化改造，从而显著提高了其资产管理和量化业务部门的投资风险和衍生品分析能力。

# 场景

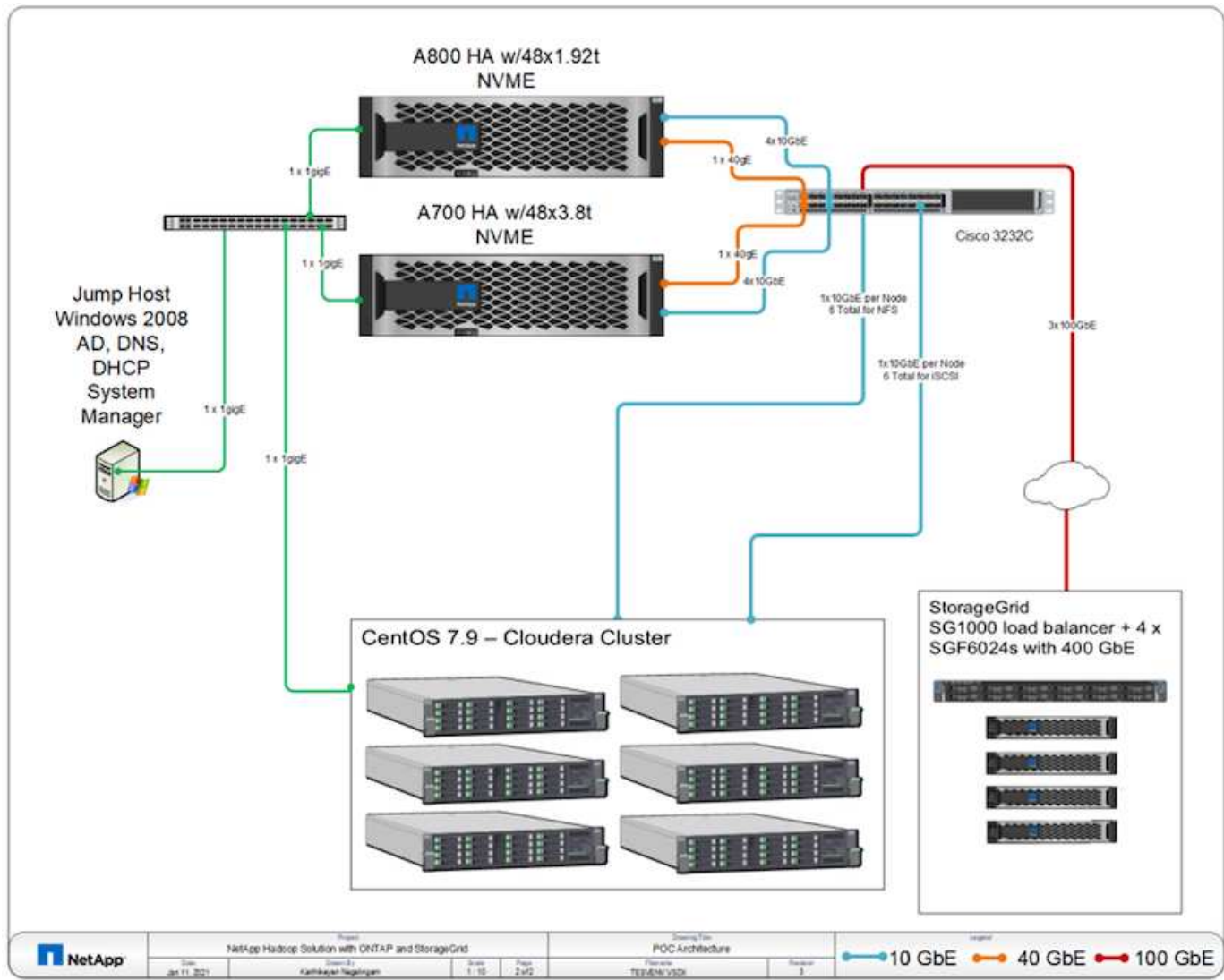
在客户现有的环境中，用于分析平台的 Hadoop 基础架构利用了 Hadoop 服务器的内部存储。由于 JBOD 环境的专有性，组织内的许多内部客户无法利用他们的蒙特卡罗定量模型，该模型是一种依赖于实时数据重复样本的模拟。对市场走势不确定性的影响的理解能力不足，对量化资产管理业务部门不利。

要求和挑战

该银行的定量业务部门需要一种有效的预测方法来实现准确、及时的预测。为此，团队认识到需要实现基础设施现代化，减少现有的 I/O 等待时间，并提高 Hadoop 和 Spark 等分析应用程序的性能，以有效模拟投资模型、衡量潜在收益和分析风险。

解决方案

客户现有的 Spark 解决方案已配备 JBOD。然后利用NetApp ONTAP、NetApp StorageGRID和 MiniIO Gateway to NFS 来减少银行量化金融小组的 I/O 等待时间，该小组对评估潜在收益和风险的投资模型进行模拟和分析。此图显示了采用NetApp存储的 Spark 解决方案。



如上图所示，部署了AFF A800、A700 系统和StorageGRID ，以便在六节点 Hadoop 集群中通过 NFS 和 S3 协议访问 parquet 文件，并使用 Spark、YARN 和 Hive 元数据服务进行数据分析操作。

客户旧环境中的直接连接存储 (DAS) 解决方案的缺点是无法独立扩展计算和存储。借助NetApp ONTAP Spark 解决方案，该银行的财务分析业务部门能够将存储与计算分离，并根据需要更有效地无缝地提供基础设施资源。

通过使用带有 NFS 的 ONTAP ，计算服务器 CPU 几乎完全用于 Spark SQL 作业，并且 I/O 等待时间减少了近 70%，从而为 Spark 工作负载提供了更好的计算能力和性能提升。随后，提高 CPU 利用率还使客户能够利用

GPU（例如 GPUDirect）进一步实现平台现代化。此外，StorageGRID为 Spark 工作负载提供了低成本的存储选项，而 MinIO Gateway 通过 S3 协议提供对 NFS 数据的安全访问。对于云中的数据，NetApp推荐使用Cloud Volumes ONTAP、Azure NetApp Files和Google Cloud NetApp Volumes。

## 结束语

本节概述了NetApp为满足各种 Hadoop 数据保护要求而提供的用例和解决方案。通过使用NetApp提供支持的数据结构，客户可以：

- 利用 NetApp 丰富的数据管理功能以及与 Hadoop 原生工作流的集成，可以灵活地选择正确的数据保护解决方案。
- 将其 Hadoop 集群备份窗口时间减少近 70%。
- 消除 Hadoop 集群备份造成的任何性能影响。
- 同时为单一分析数据源提供来自不同云提供商的多云数据保护和数据访问。
- 使用FlexClone技术创建快速且节省空间的 Hadoop 集群副本。

## 在哪里可以找到更多信息

要了解有关本文档中描述的信息的更多信息，请参阅以下文档和/或网站：

- NetApp大数据分析解决方案

["https://www.netapp.com/us/solutions/applications/big-data-analytics/index.aspx"](https://www.netapp.com/us/solutions/applications/big-data-analytics/index.aspx)

- 使用NetApp存储的 Apache Spark 工作负载

<https://www.netapp.com/pdf.html?item=/media/26877-nva-1157-deploy.pdf>

- 适用于 Apache Spark 的NetApp存储解决方案

["https://www.netapp.com/media/16864-tr-4570.pdf"](https://www.netapp.com/media/16864-tr-4570.pdf)

- NetApp支持的数据结构上的 Apache Hadoop

["https://www.netapp.com/media/16877-tr-4529.pdf"](https://www.netapp.com/media/16877-tr-4529.pdf)

## 致谢

- Paul Burland，NetApp澳新银行维多利亚区销售部销售代表
- NetApp业务发展经理 Hoseb Dermanilian
- NetApp MPSG 总监 Lee Dorrier
- NetApp澳新银行维多利亚区 SE 系统工程师 David Thiessen

## 版本历史

| 版本    | 日期       | 文档版本历史           |
|-------|----------|------------------|
| 1.0 版 | 2018年1月  | 初始版本             |
| 2.0 版 | 2021年10月 | 更新用例 #5：加速分析工作负载 |
| 3.0 版 | 2023年11月 | 删除了 NIPAM 详细信息   |

## 版权信息

版权所有 © 2025 NetApp, Inc.。保留所有权利。中国印刷。未经版权所有者事先书面许可，本文档中受版权保护的任何部分不得以任何形式或通过任何手段（图片、电子或机械方式，包括影印、录音、录像或存储在电子检索系统中）进行复制。

从受版权保护的 NetApp 资料派生的软件受以下许可和免责声明的约束：

本软件由 NetApp 按“原样”提供，不含任何明示或暗示担保，包括但不限于适销性以及针对特定用途的适用性的隐含担保，特此声明不承担任何责任。在任何情况下，对于因使用本软件而以任何方式造成的任何直接性、间接性、偶然性、特殊性、惩罚性或后果性损失（包括但不限于购买替代商品或服务；使用、数据或利润方面的损失；或者业务中断），无论原因如何以及基于何种责任理论，无论出于合同、严格责任或侵权行为（包括疏忽或其他行为），NetApp 均不承担责任，即使已被告知存在上述损失的可能性。

NetApp 保留在不另行通知的情况下随时对本文档所述的任何产品进行更改的权利。除非 NetApp 以书面形式明确同意，否则 NetApp 不承担因使用本文档所述产品而产生的任何责任或义务。使用或购买本产品不表示获得 NetApp 的任何专利权、商标权或任何其他知识产权许可。

本手册中描述的产品可能受一项或多项美国专利、外国专利或正在申请的专利的保护。

有限权利说明：政府使用、复制或公开本文档受 DFARS 252.227-7013（2014 年 2 月）和 FAR 52.227-19（2007 年 12 月）中“技术数据权利 — 非商用”条款第 (b)(3) 条规定的限制条件的约束。

本文档中所含数据与商业产品和/或商业服务（定义见 FAR 2.101）相关，属于 NetApp, Inc. 的专有信息。根据本协议提供的所有 NetApp 技术数据和计算机软件具有商业性质，并完全由私人出资开发。美国政府对这些数据的使用权具有非排他性、全球性、受限且不可撤销的许可，该许可既不可转让，也不可再许可，但仅限在与交付数据所依据的美国政府合同有关且受合同支持的情况下使用。除本文档规定的情形外，未经 NetApp, Inc. 事先书面批准，不得使用、披露、复制、修改、操作或显示这些数据。美国政府对国防部的授权仅限于 DFARS 的第 252.227-7015(b)（2014 年 2 月）条款中明确的权利。

## 商标信息

NetApp、NetApp 标识和 <http://www.netapp.com/TM> 上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。