技術報告

How to enable StorageGRID in your environment

NetApp
April 26, 2024

目錄

技術報告

NetApp StorageGRID 與巨量資料分析

NetApp StorageGRID 使用案例

NetApp StorageGRID 物件儲存解決方案提供擴充性、資料可用度、安全性及高效能。各種規模的組織、以及各行各業的組織、都會使用 StorageGRID S3 來處理各種使用案例。讓我們來探索一些典型案例：

- 巨量資料分析： * StorageGRID S3 經常用作資料湖、企業可在其中儲存大量結構化和非結構化資料、以便使用 Apache Spark 、 Splunk Smartstore 和 Dremio 等工具進行分析。
- 資料分層： * NetApp 客戶使用 ONTAP 的 FabricPool 功能、在高效能本機層之間自動將資料移至 StorageGRID 。分層可釋放昂貴的 Flash 儲存空間、以儲存熱資料、同時在低成本物件儲存設備上隨時提供冷資料。如此可將效能與節約效益發揮到極致。
- 資料備份與災難恢復： * 企業可以使用 StorageGRID S3 做為可靠且具成本效益的解決方案、在發生災難時備份關鍵資料並加以恢復。
- 應用程式的資料儲存： * StorageGRID S3 可作為應用程式的儲存後端、讓開發人員輕鬆儲存及擷取檔案、影像、影片及其他類型的資料。
- 內容交付： * StorageGRID S3 可用於儲存靜態網站內容、媒體檔案及軟體下載、並提供給全球各地的使用者、運用 StorageGRID 的地理發佈和全球命名空間、提供快速可靠的內容交付。
- 資料分層： * NetApp 客戶使用 ONTAP FabricPool 功能、在高效能本機層之間自動將資料移至 StorageGRID 。分層可釋放昂貴的 Flash 儲存空間、以儲存熱資料、同時讓低成本物件儲存設備隨時可用冷資料。如此可將效能與節約效益發揮到極致。
- 資料歸檔： * StorageGRID 提供不同的儲存類型、並支援分層處理至公有長期低成本儲存選項、因此是歸檔及長期保留資料的理想解決方案、這些資料必須保留以供法規遵循或歷史用途。
- 物件儲存使用案例 *

[StorageGRID 使用案例圖表、寬度 = 396 、高度 = 394]

在上述情況中、巨量資料分析是最熱門的使用案例之一、其使用率也逐漸上升。

為何選擇 StorageGRID 來處理資料湖？

- 更高的協同作業效率：大規模共用多站台、多租戶、並具備業界標準 API 存取功能
- 降低營運成本：單一、自我修復、自動化橫向擴充架構的操作簡易性
- 擴充性：與傳統的 Hadoop 和資料倉儲解決方案不同、 StorageGRID S3 物件儲存設備可將儲存設備與運算和資料分離、讓企業能夠隨著成長擴充儲存需求。
- 耐用性與可靠性： StorageGRID 提供 99.999% 的耐用度、這表示儲存的資料對於資料遺失具有高度抵抗能力。它也提供高可用度、確保資料隨時可供存取。
- 安全性： StorageGRID 提供各種安全功能、包括加密、存取控制原則、資料生命週期管理、物件鎖定和版本設定、以保護儲存在 S3 儲存區中的資料
- StorageGRID S3 資料湖 *

[StorageGRID datalake 範例、 width=614 、 height =345]

哪一個資料倉儲或資料湖最適合搭配 S3 物件儲存設備使用

NetApp 以三種資料倉儲 / 湖屋生態系統（ Hive 、 Delta Lake 和 Dremio ）為基準測試 StorageGRID 。 "Apache 冰山：最終指南" 包括資料倉儲和資料湖屋的簡介、以及這兩種架構的優缺點。

- 基準測試工具 - TPC-DS - https://www.tpc.org/tpcds/
- Big Data 生態系統
  - 由 5 個 VM 叢集所構成、每個 VM 都有 128G RAM 和 24 個 vCPU 、以及用於系統磁碟的 SSD 儲存設備
  - Hadoop 3.3.5 搭配 Hive 3.1.3 （ 1 個名稱節點 + 4 個資料節點）
  - Delta Lake with Spark 3.2.0 （ 1 位大師 + 4 位員工）和 Hadoop 3.3.5
  - Dremio V23 （ 1 位碩士 + 4 位執行者）
- 物件儲存
  - NetApp ^ ® ^ StorageGRID ^ ® ^ 11.6 含 3 個 SG6060 + 1 個 SG1000 負載平衡器
  - 物件保護 - 2 份複本
- 資料庫大小 1000GB
- 所有 3 個生態系統都停用快取、以取得每項查詢測試的一致結果。

TPC-DS 隨附 99 個複雜的 SQL 查詢、可用於查詢基準測試。我們測量了完成所有 99 項查詢所需的總分鐘數、並將 S3 要求的類型和數量逐一列出、以進一步分析結果。下表顯示所有 99 個查詢的總持續時間、第二個表格則摘要說明每個傳送至 StorageGRID 的生態系統中 S3 要求的數量和類型。

- TPC-DS 查詢結果 *

生態系統	**Hive**	德爾塔湖	夢中
儲存層	NetApp ^ ® ^ StorageGRID ^ ® ^	NetApp ^ ® ^ StorageGRID ^ ® ^	NetApp ^ ® ^ StorageGRID ^ ® ^
磁碟機類型	HDD	HDD	HDD
表格格式	硬地板	硬地板	硬地板 [1]
資料庫大小	1000G	1000G	1000G
TPCDS 99 查詢 總分鐘數	1084 [2]	55	47

[1] 測試了 Parquet 和 iceberg 表格格式、結果類似。

[2] Hive 無法完成查詢編號 72 。

- TPC-DS 查詢 - S3 要求明細 *

S3 要求	**Hive**	德爾塔湖	夢中
取得	1,117,184	2,074,610	4,414,227

S3 要求	Hive	德爾塔湖	夢中
觀察： 全方位實現	從 32 MB 物件到 2 KB 到 2 MB 的範圍達到 80% 、每秒 50 到 100 個要求	73% 的範圍從 32 MB 物件低於 100KB 、從 1000 到 1400 個要求 / 秒	從 256 MB 物件獲得 90% 的 100 萬位元組範圍、2000 年至 2300 個要求 / 秒
列出物件	312,053	24 、 158	240
標題 （不存在的物件）	156,027	12 、 103	192.
標題 （存在的物件）	982,126.	922,732.	1845
申請總數	2,567,390	3 、 033 、 603	4,416504..

從第一張表格中、我們可以看到 Delta Lake 和 Dremio 比 Hive 快得多。從第二個表格中、我們注意到 Hive 傳送了許多 S3 清單物件要求、這在所有物件儲存平台中通常都很慢、尤其是在處理包含許多物件的儲存區時。如此可大幅增加整體查詢持續時間。另一項觀察是 Dremio 能夠同時傳送大量的 GET 要求、每秒 2 、 000 至 2 、 300 個要求、而 Hive 則是每秒 50 至 100 個要求。Hive 和 Hadoop S3A 模擬標準檔案系統、有助於 S3 物件儲存作業緩慢。

搭配 Hive 或 Spark 使用 Hadoop （在 HDFS 或 S3 物件儲存設備上）需要對 Hadoop 和 Hive/Spark 有廣泛的瞭解、以及每項服務的設定如何互動、而且它們有 1000 多種設定。通常、這些設定是相互關聯的、無法單獨變更。要找到最佳的設定和值組合、需要花費大量的時間和精力。

Dremio 是資料湖引擎、使用端點對端 Apache Arrow 來大幅提升查詢效能。Apache Arrow 提供標準化的列式記憶體格式、可實現高效率的資料共享和快速分析。Arrow 採用語言不相關的方法、旨在免除資料序列化和反序列化的需求、改善複雜資料程序和系統之間的效能和互通性。

Dremio 的效能主要是由 Dremio 叢集的運算能力所驅動。雖然 Dremio 使用 Hadoop 的 S3A 連接器進行 S3 物件儲存連線、但 Hadoop 並不需要、而 Dremio 也不使用 Hadoop 的 FS.s3a 設定。如此一來、無需花時間學習和測試各種 Hadoop s3a 設定、即可輕鬆調整 Dremio 效能。

從這個基準測試結果中、我們可以得出結論、針對 S3 型工作負載最佳化的大型資料分析系統是主要的效能因素。Dremio 可最佳化查詢執行、有效運用中繼資料、並提供對 S3 資料的無縫存取、因此相較於使用 S3 儲存設備時的 Hive 、效能更佳。請參閱此 "頁面" 使用 StorageGRID 設定 Dremio S3 資料來源。

請造訪下列連結、深入瞭解 StorageGRID 和 Dremio 如何合作提供現代化且有效率的資料湖基礎架構、以及 NetApp 如何從 Hive + HDFS 移轉至 Dremio + StorageGRID 、大幅提升巨量資料分析效率。

- "利用 NetApp StorageGRID 大幅提升巨量資料的效能"
- "StorageGRID 和 Dremio 提供現代化、功能強大且有效率的資料湖基礎架構"
- "NetApp 如何透過產品分析重新定義客戶體驗"

Hadoop S3A 調校

Hadoop S3A 連接器可促進 Hadoop 應用程式與 S3 物件儲存之間的順暢互動。調整 Hadoop S3A Connector 是在使用 S3 物件儲存設備時最佳化效能的關鍵。在我們開始調整詳細資料之前、讓我們先基本瞭解 Hadoop 及其元件。

什麼是 **Hadoop** ？

- Hadoop * 是功能強大的開放原始碼架構、專為處理大規模資料處理與儲存而設計。它可跨電腦叢集進行分散式儲存和平行處理。

Hadoop 的三個核心元件為：

- * Hadoop HDFS （ Hadoop 分散式檔案系統） * ：此功能可處理儲存、將資料分成區塊、並在節點之間散佈。
- * Hadoop MapReduce * ：負責處理資料、將工作分割成較小的區塊、並平行執行。
- * 哈大諾布（又是另一個資源協商者）： * "有效管理資源並排程工作"

Hadoop HDFS 和 **S3A** 接頭

HDFS 是 Hadoop 生態系統的重要元件、在高效率的巨量資料處理中扮演重要角色。HDFS 提供可靠的儲存與管理功能。它可確保平行處理和最佳化的資料儲存、進而加快資料存取和分析速度。

在巨量資料處理中、 HDFS 在為大型資料集提供容錯儲存方面表現優異。它透過資料複寫來達成此目標。它可以在資料倉儲環境中儲存及管理大量的結構化和非結構化資料。此外、它還能與領先業界的巨量資料處理架構（例如 Apache Spark 、 Hive 、 Pig 和 Fllink ）無縫整合、實現可擴充且有效率的資料處理。它與 Unix （ Linux ）作業系統相容、是偏好使用 Linux 環境進行巨量資料處理的組織的理想選擇。

隨著資料量隨著時間成長、將新機器新增至 Hadoop 叢集、並使用其本身的運算和儲存設備的方法變得效率不彰。線性擴充會在有效使用資源和管理基礎架構方面帶來挑戰。

為了因應這些挑戰、 Hadoop S3A 連接器提供高效能 I/O 、可與 S3 物件儲存設備相較。運用S3A實作Hadoop工作流程、有助於將物件儲存設備當作資料儲存庫、並可將運算與儲存設備分開、進而獨立擴充運算與儲存設備。分離運算和儲存設備也能讓您將適當的資源量用於運算工作、並根據資料集的大小提供容量。因此、您可以降低Hadoop工作流程的整體TCO。

Hadoop S3A 接頭調校

S3 的行為與 HDFS 不同、有些為了保留檔案系統外觀的嘗試則不佳。為了最有效地使用 S3 資源、必須仔細調整 / 測試 / 實驗。

本文中的 Hadoop 選項是以 Hadoop 3.3.5 為基礎、請參閱 "Hadoop 3.3.5 core-site.xml" 適用於所有可用選項。

附註：在每個 Hadoop 版本中、某些 Hadoop FS.s3a 設定的預設值會有所不同。請務必查看您目前 Hadoop 版本的預設值。如果這些設定未在 Hadoop 核心網站 .xml 中指定、則會使用預設值。您可以使用 Spark 或 Hive 組態選項、在執行階段覆寫值。

您一定要來這裏 "Apache Hadoop 頁面" 以瞭解每個 FS.s3a 選項。如有可能、請在非正式作業 Hadoop 叢集中測試它們、以找出最佳值。

您應該閱讀 "使用 S3A 連接器時發揮最大效能" 以取得其他調整建議。

讓我們來探討一些重要考量：

- 1。資料壓縮 *

請勿啟用 StorageGRID 壓縮。大部分的巨量資料系統使用位元組範圍 GET 、而非擷取整個物件。使用位元組

範圍取得壓縮物件會大幅降低取得效能。

2.S3A 交付人 *

一般而言、建議使用 magic s3a committer 。請參閱此 "一般 S3A 交付選項頁面" 更深入瞭解魔術棒及其相關的 s3a 設定。

Magic Committer ：

Magic Committer 特別仰賴 S3Guard 在 S3 物件存放區上提供一致的目錄清單。

有了一致的 S3 （現在就是這樣）、萬智牌委員會就能安全地與任何 S3 儲存桶搭配使用。

選擇與實驗：

視您的使用案例而定、您可以在 Staging Committer （依賴叢集 HDFS 檔案系統）和 Magic Committer 之間進行選擇。

請嘗試兩者、以判斷哪一種最適合您的工作負載和需求。

總結來說、 S3A Commiters 提供解決方案、以因應對 S3 的一致、高效能及可靠輸出承諾這項基本挑戰。其內部設計可確保資料傳輸效率、同時維持資料完整性。

[S3A 選項表]

- 3.執行緒、連線集區大小和區塊大小 *
- 每個與單一貯體互動的 **S3A** 用戶端都有其專屬的開放式 HTTP 1.1 連線集區和執行緒、可用於上傳及複製作業。
- "您可以調整這些集區大小、以在效能與記憶體 / 執行緒使用量之間取得平衡"。
- 將資料上傳至 S3 時、資料會分成區塊。預設區塊大小為 32 MB 。您可以設定 FS.s3a.block.size 屬性來自訂此值。
- 較大的區塊大小可降低上傳期間管理多個零件的成本、進而改善大型資料上傳的效能。大型資料集的建議值為 256 MB 或更高。

[S3A 選項表]

*4.多部分上傳 *

s3a commiters * Always * 使用 MPU （多部分上傳）將資料上傳至 S3 儲存庫。這是為了允許：工作失敗、工作的投機性執行、以及工作在提交前中止。以下是與多部分上傳相關的一些重要規格：

- 最大物件大小： 5 TiB （ TB ）。
- 每次上傳的最大零件數： 10 、 000 。
- 零件編號：範圍從 1 到 10 、 000 （含）。
- 零件大小：介於 5 MIB 和 5 GiB 之間。值得注意的是、您的多部分上傳最後一部分沒有最小大小限制。

將較小的零件大小用於 S3 多個部分上傳、既有優點也有缺點。

- 優點 * ：

- 從網路快速恢復問題：當您上傳較小的零件時、因為網路錯誤而重新啟動失敗上傳的影響會降至最低。如果零件失敗、您只需要重新上傳該特定零件、而不需要重新上傳整個物件。
- 更佳的平行化：可以同時上傳更多零件、同時利用多執行緒或並行連線。這種平行化可提升效能、尤其是處理大型檔案時。
- 缺點 * ：
- 網路負荷：較小的零件尺寸代表需要上傳更多零件、每個零件都需要自己的 HTTP 要求。更多 HTTP 要求會增加啟動和完成個別要求的成本。管理大量的小型零件可能會影響效能。
- 複雜度：管理訂單、追蹤零件、以及確保成功上傳可能會很麻煩。如果需要中止上傳、則需要追蹤和清除所有已上傳的零件。

若為 Hadoop 、建議使用 256 MB 或以上的零件大小、以使用 FS.s3a.multite.size 。請務必將 FS.s3a.mutlipart.threshold 值設為 2 x FS.s3a.multipe.size 值。例如、 fs.s3a.multipe.size = 256M 、 fs.s3a.mutlipart.threshold 應為 512M 。

大型資料集使用較大的零件大小。請務必根據您的特定使用案例和網路條件、選擇零件尺寸來平衡這些因素。

多部分上傳是 "三步驟程序"：

1. 上傳即會啟動、 StorageGRID 會傳回上傳 ID 。
2. 物件零件會使用 upload-id.
3. 上傳所有物件零件後、會傳送完整的多個部分上傳要求與 upload-id.StorageGRID 會從上傳的零件建構物件、用戶端可以存取物件。

如果未成功傳送完整的多部分上傳要求、則零件會留在 StorageGRID 中、不會建立任何物件。當工作中斷、失敗或中止時、就會發生這種情況。零件會保留在網格中、直到多個零件上傳完成或中止、或 StorageGRID 在上傳開始後 15 天內清除這些零件。如果在某個儲存庫中有許多（數十萬到數百萬）進行中的多部分上傳、當 Hadoop 傳送「 list-multify-upload 」（此要求不依上傳 ID 篩選）時、要求可能需要很長時間才能完成或最終逾時。您可以考慮將 FS.s3a.mutlipart.purge 設為 true 、並設定適當的 FS.s3a.multipe.pure.age 值（例如 5 至 7 天、請勿使用 86400 的預設值、即 1 天）。或請 NetApp 支援人員調查情況。

[S3A 選項表]

*5.緩衝區將資料寫入記憶體 *

若要提升效能、您可以在將資料上傳至 S3 之前、先緩衝寫入記憶體中的資料。這樣可以減少小寫入次數、並提高效率。

[S3A 選項表]

請記住、 S3 和 HDFS 的運作方式各不相同。為了最有效地使用 S3 資源、必須仔細調整 / 測試 / 實驗。