IBM SPSS Analytic Server
3.2.1 版

管理手冊

IBM

附註

在使用本資訊及它支援的產品之前，請閱讀第 21 頁的『注意事項』中的資訊。

產品資訊

此版本適用於 IBM SPSS Analytic Server 3.2.1 版及所有後續版本與修訂，但新版本另有說明者不在此限。

目錄

第 1 章 承租人管理

承租人提供使用者、專案及資料來源的高層次分割，從而無法在承租人之間共用物件。 每一個使用者都可存取其指派之承租人環境定義中的系統。

您可以在 Analytic Server 主控台中，管理承租人，以及將使用者指派給承租人。 「承租人」頁面的視圖取決於登入至主控台之使用者的角色：

- 安裝期間設定的「超級使用者」管理者是承租人管理員。 只有此使用者可以建立新的承租人，以及編輯任何承租人的內容。
- 具有「管理者」角色的使用者可以編輯他們登入之承租人的內容。
- 具有「使用者」角色的使用者無法編輯承租人內容。 「承租人」頁面對他們隱藏。
- 具有讀者角色的使用者不能編輯資料來源，甚至無法登入 Analytic Server 主控台。

管理者可以存取「專案」及「資料來源」頁面，以及管理任何專案或資料來源以進行清理及管理。 如需相關資訊，請參閱 *IBM® SPSS® Analytic Server* 使用手冊。

承租人清單

主要「承租人」頁面在表格中顯示現有承租人。 只有「超級使用者」管理者能夠在此頁面上進行編輯。

- 按一下承租人名稱，以顯示其詳細資料，並編輯其內容。
- 按一下承租人 URL，以在該承租人的環境定義中開啟主控台。

  註：您將登出主控台，並將需要使用承租人的有效認證來登入。
- 在搜尋區中鍵入內容可過濾清單，以僅顯示名稱中含有搜尋字串的承租人。
- 按一下**新建**，以使用您在**新增承租人**對話框中指定的名稱建立新承租人。 請參閱第 2 頁的『命名規則』，以取得您可以為承租人提供之名稱的限制。
- 按一下**刪除**，以移除選取的承租人。
- 按一下**重新整理**，以更新清單。

個別承租人詳細資料

內容區劃分為數個可收合區段。

詳細資料

名稱 一個可編輯的文字欄位，顯示承租人的名稱。

說明 一個可編輯的文字欄位，容許您提供關於租戶的解釋性文字。

URL 此 URL 提供給使用者，以透過 Analytic Server 主控台登入承租人，以及用來配置 SPSS Modeler 伺服器。 如需配置 SPSS Modeler 的詳細資料，請參閱 *IBM SPSS Analytic Server Installation and Configuration Guide*。

狀態 **作用中**承租人目前正在使用中。 讓承租人處於**非作用中Inactive**會阻止使用者登入該承租人，但不會刪除任何基礎資訊。

主體

主體是從安裝期間設定之安全提供者處得來的使用者和群組。 您可以將主體作為「管理者」、「使用者」或「讀者」新增至租戶。

- 在文字框中輸入內容可過濾名稱中含有搜尋字串的使用者和群組。 從下拉清單中選取**管理者**、**使用者**或**讀者**，以指派其在租戶內的角色。 按一下**新增參與者**，將他們新增至作者清單。
- 若要移除參與者，請在成員清單中選取使用者或群組，然後按一下**移除參與者**。

度量值

可讓您配置租戶的資源限制。 報告租戶目前使用的磁碟空間。

- 您可以設定租戶的磁碟空間配額上限；當達到此限制時，無法將更多資料寫入至此租戶上的磁碟，除非清除足夠的磁碟空間，讓租戶磁碟空間用量低於配額。
- 您可以設定租戶的磁碟空間警告層次；當超出該配額時，此租戶上的主體無法提交任何分析工作，除非清除足夠的磁碟空間，讓租戶磁碟空間用量低於配額。
- 您可以設定此租戶上於單一時間可以執行的平行工作數目上限；當超出該配額時，此租戶上的主體無法提交任何分析工作，除非目前執行中的工作完成。
- 您可以設定資料來源可以具有的欄位數目上限。 每當建立或更新資料來源時，會檢查該限制。
- 您可以設定檔案大小上限 (MB)。 上傳檔案時，會檢查該限制。

安全提供者配置

可讓您指定使用者鑑別提供者。 **預設值**使用在安裝及配置期間設定的預設承租人提供者。 **LDAP** 可讓您向外部 LDAP 伺服器（例如 Active Directory 或 OpenLDAP）鑑別使用者。 指定提供者的設定，並選擇性地指定過濾器設定，以控制「主體」區段中可用的使用者和群組。

命名規則

對於 Analytic Server 中可以給定唯一名稱的任何項目（如資料來源及專案），下列規則適用於那些名稱。

- 在單個租戶內，名稱在相同類型的物件內必須是唯一的。 例如，兩個資料來源無法同時命名為 insuranceClaims，但一個資料來源及一個專案可以分別命名為 insuranceClaims。
- 名稱區分大小寫。 例如，insuranceClaims 與 InsuranceClaims 會視為唯一名稱。
- 名稱忽略前導及尾端空格。
- 下列字元在名稱中無效。

  `~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n`

第 2 章 使用者入門

告知使用者導覽至 `http://<host>:<port>/<context-root>/admin/<tenant>`，並輸入其使用者名稱與密碼，從而登入至 Analytic Server 主控台。

註：在 Analytic Server 主控台登入提示期間輸入的使用者名稱是沒有領域名稱字尾的。 結果，在定義多個領域時，將向使用者呈現**領域**下拉清單，讓他們能夠選取適當的領域。 若僅定義一個領域，則當使用者登入 Analytic Server 時，不會呈現**領域**下拉清單。

<host>
Analytic Server 主機的位址。

<port>
Analytic Server 接聽的埠。 依預設，這是 9080。

<context-root>
Analytic Server 的環境定義根目錄。 依預設，這是分析伺服器。

<tenant>
在多租戶環境中，表示您所屬的租戶。 在單一租戶環境中，預設租戶為 **ibm**。

例如，如果主機的 IP 位址為 9.86.44.232，您已建立租戶 "mycompany" 並向其新增了使用者，且將其他設定保留為其預設值，則使用者應該導覽至 `http://9.86.44.232:9080/analyticserver/admin/mycompany`，以存取 Analytic Server 主控台。

第 3 章 Analytic Server 工作名稱

Analytic Server 產生對映化簡工作及 Spark 工作，可以透過 Hadoop 叢集的「資源管理程式」使用者介面進行監視。

對映化簡工作具有下列結構。

```
AS/{tenant name}/{user name}/{algorithm name}
```

{tenant name}
這是執行工作所使用的租戶名稱。

{user name}
這是要求工作的使用者。

{algorithm name}
這是工作中的主要演算法。 請注意，單一串流可能產生多個對映化簡工作；類似地，串流內的數個作業可以包含在單一對映化簡工作內。

「資源管理程式」使用者介面中會顯示所有對映化簡工作。 單一 Spark 應用程式針對每一個 Analytic Server 啟動。 開啟 Spark 應用程式的使用者介面可監視 Spark 工作（**說明**直欄中會顯示工作名稱）。

第 4 章 IBM SPSS Analytic Server 最佳實務及建議

下列區段提供有關資料資源、叢集配置及 IBM SPSS Modeler 串流的 Analytic Server 最佳實務及建議。

資料來源

Analytic Server 支援下列資料來源類型：

- 檔案型資料來源，例如定界文字、固定文字及 Microsoft Excel 檔案。
- 關聯式資料庫，例如 DB2、Oracle、Microsoft SQL Server、Teradata、Postgres、Netezza、MySQL 及 Amazon Redshift。
- 包括所有內建資料類型（例如 ORC 及 Parquet）的 Hive/HCatalog 資料來源，以及適用 Hive Serializer-Deserializer 實作可用的任何自訂資料類型。 此外，可以配置 Analytic Server 以存取 NoSQL 資料庫，例如 HBase、MongoDB、Accumulo、Cassandra、Oracle NoSQL，以及適當 Hive Storage Handler 實作可用的其他資料庫。
- 地理空間類型資料來源（Shape 檔型與對映服務型）。

Hive/HCatalog 資料來源的 Analytic Server 限制

- 如果 SPSS Modeler 的「選取」節點需要 Hive 推回，則過濾表示式只能參照類型為 STRING 的分割直欄。從 Analytic Server 3.0 開始，下列分割直欄已新增資料類型支援：TINYINT、SMALLINT、INT、BIGINT。為 Hive 資料來源指定的靜態過濾表示式可能具有任何資料類型之分割直欄的過濾表示式。
- Analytic Server 不支援基於 Hive 視圖的資料來源。

叢集配置 - 安全

Kerberos 模擬

在 3.0.1 版之前，Analytic Server 實例利用 Analytic Server 金鑰表中的「使用者主體名稱」，以在啟用 Kerberos 安全時鑑別 HDFS 作業。 從 3.0.1 版開始，Analytic Server 利用 Analytic Server 金鑰表中的「服務主體名稱」，以及要求的使用者名稱（屬於發出剩餘要求的使用者），以鑑別利用 Kerberos 模擬的 HDFS 作業。 需要 Analytic Server 3.0.1 或更新版本，以在 Kerberos 啟用的叢集中執行時，將模擬配置屬性新增至 HDFS（或者 Hive 服務配置）。 在使用 HDFS 時，必須將下列內容新增至 HDFS `core-site.xml` 檔：

```
hadoop.proxyuser.<analytic_server_service_principal_name> .hosts = *
hadoop.proxyuser.<analytic_server_service_principal_name> .groups = *
```

其中，`<analytic_server_service_principal_name>` 是在 Analytic Server 配置的 `Analytic_Server_User` 欄位中指定的預設 `as_user` 值。

如果透過 Hive/HCatalog 從 HDFS 存取資料，則下列內容也必須新增至 HDFS `core-site.xml` 檔：

```
hadoop.proxyuser.hive.hosts = *
hadoop.proxyuser.hive.groups = *
```

Kerberos 跨領域鑑別

Analytic Server 支援 Kerberos 跨領域鑑別。 若要啟用此特性，您必須首先確保 KDC 跨領域鑑別已啟用，然後將下列設定新增至 Analytic Server Ambari 配置的 **Custom analytics.cfg** 區段：

```
kerberos.user.realm.trim = true
```

叢集配置 - 效能調整設定及結果

Spark 配置

Analytic Server 使用 yarn-client 模式與 YARN 互動，並在 Hadoop 叢集上執行 Spark 工作。

Analytic Server 自訂配置：

- Ambari 設定在 Analytic Server Ambari 配置的**自訂 analytics.cfg** 區段中定義。
- Cloudera 設定位於 Cloudera Manager 之 **analyticserver-conf/config.properties** 的 **Analytic Server 進階配置 Snippet（安全閥）**區段中。

1. 考量增大 **spark.driver.memory** 配置設定的值，方法是在 Analytic Server 自訂配置中新增配置項目（若未明確設定，則預設值為 1g）。 例如：

   ```
   spark.driver.memory=2g
   ```

2. 選取下列其中一個使用 Spark 之 Analytic Server 的資源使用選項。
   - **選項 A：靜態資源配置的配置**

     必須在 Analytic Server 自訂配置中配置如下 3 個參數：

     ```
     spark.executor.instances
     spark.executor.cores
     spark.executor.memory
     ```

     下列步驟說明如何判定參數值。

     a. 以 CPU 及記憶體形式，建立 Analytic Server 可以永久地為 Spark 配置的百分比。 這會產生特定的核心數目 (C) 及可以在每一個機器中使用的固定記憶體數量 (M)。

     b. 建立每一個機器可以執行的執行程式數目 (E)。 這些執行程式在每一個叢集節點上作為單獨的 Hadoop 儲存器（處理程序）執行。 通常，大於 2 的值適當，但是該值必須小於核心總數。 為 Spark 配置的記憶體會在這些執行程式之間分割，因此為此參數選取較高值會降低為每一個儲存器配置的記憶體數量。

     c. 建立每一個執行程式使用的核心數目 (E)。 通常，此值是 C/E（每一個機器為 Spark 應用程式配置的核心數目，除以執行程式總數）。

     d. 建立用於每一個執行程式的記憶體數量 (ME)。 這通常是 M/E。

     註：使用的執行程式及核心數目必須平衡，即每一個執行程式記憶體的數量應該大於 3G * CE。 每一個執行程式的每一個核心必須至少配置有 3G 記憶體用作儲存體或計算記憶體。

     ```
     spark.executor.instances = <E>*N /<E> // value established in step b where N is the number of compute nodes
     spark.executor.cores = <CE> // value established in step c
     spark.executor.memory = <ME> // value established in step d
     ```

圖 *1.* 自訂 *analytics.cfg Spark* 設定

   - **選項 B：動態資源配置的配置**

     如果使用此選項，則會根據整個叢集的實際可用資源動態增加/減少由 YARN 配置的所有執行程式。

最低配置為：

```
spark.dynamicAllocation.enabled = true
spark.shuffle.service.enabled = true
```

一般配置為：

```
spark.default.emitter.class = com.spss.ae.spark.ListEmitter
spark.default.emitter.compressed = false
spark.dynamicAllocation.enabled = true
spark.executor.cores = 4
spark.executor.memory = 16g
spark.io.compression.codec = snappy
spark.rdd.compress = true
spark.shuffle.service.enabled = true
```

附註：

- 不應使用 `spark.executor.instances = <E>`，否則將採用靜態資源配置。
- 關於執行程式核心和記憶體值的考量選項 A 相同。

3. 您可以使用下列設定停用 Analytic Server 自訂配置中的 Spark 快取：

```
spark.cache=false
spark.storage.memoryFraction = 0.3
```

圖 2. 自訂 *analytics.cfg Spark* 快取設定

使用大型 IBM SPSS Modeler 串流時，不得停用 Spark 快取。 在此實例中停用 Spark 快取會導致執行中串流更緩慢，但是會避免每個執行程式的記憶體數量較小時可能發生的記憶體不足狀況。

JVM 配置

Ambari 設定：

1. 在 Analytic Server Ambari 配置中，設定伺服器可以用於本端處理的記憶體數量。 預設值 (2 GB) 可以安全地用於小型至中型串流，但是更高值的資料堆大小（例如，10 GB）應該用於更大的串流。

   分析伺服器 > 配置 > 進階 analytic-jvm-options

2. 將 `-Xmx2048M` 取代為 `-Xmx10G`，儲存配置，並重新啟動 Analytic Server。

圖 3. 進階 *analytic-jvm-options* 設定

Cloudera 設定：

1. 在 Cloudera 管理程式中，導覽至 Analytic Server 服務的**配置**標籤，並更新 `jvm-options` 控制項以設定伺服器可用來進行本端處理的記憶體量。 預設值 (2 GB) 可以安全地用於小型至中型串流，但是更高值的資料堆大小（例如，10 GB）應該用於更大的串流。

   Analytic Server 服務 > 配置 > jvm-options

2. 將 `-Xmx2048M` 取代為 `-Xmx10G`，儲存配置，並重新啟動 Analytic Server。

Yarn MapReduce2 配置：

- 如果您必須將 MapReduce 工作與 Analytic Server 的 Spark 工作平行執行，則必須將 Yarn 叢集配置為每一個 Yarn 儲存器具有至少 4 GB 記憶體。

Zookeeper 配置：

- Cloudera 需要您手動更新 Zookeeper 配置。 如需相關資訊，請參閱 https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_zookeeper_server_maintain.html。
- 如果您使用複式 SPSS Modeler 串流或大量資料（大量欄位），則可能遇到由於 Analytic Server–Zookeeper 連線岔斷而造成的工作失敗問題。 問題是 SPSS Modeler 伺服器傳送至 Analytic Server 之較大問題大小的結果。 該問題很少在 Analytic Server 3.0（或更新版本）中發生。 請使用下列步驟，以解決該問題：
  1. 在 Ambari 主控台中，導覽至 Zookeeper 服務**配置**標籤，將下列行新增至**進階 zookeeper-env** 下的 zookeeper-env 範本，然後重新啟動 Zookeeper 服務。

     ```
     export JVMFLAGS="-Xmx2048m -Djute.maxbuffer=2097152"
     ```

圖 4. *zookeeper-env* 範本設定

  2. 在 Ambari 主控台中，導覽至 Analytic Server 服務的**配置**標籤，將下列項目新增至**進階 analytics-jvm-options**，然後重新啟動 Analytic Server 服務。

     ```
     -Djute.maxbuffer=2097152
     ```

圖 5. 進階 *analytics-jvm-options* 設定

註：如果問題持續存在，請在兩個工作區中，將 -Djute.maxbuffer 值從 2097152 增加至 4194304。

IBM SPSS Modeler 串流建議

註：大部分下列建議也適用於少量資料。

少量資料上的原型

當您使用串流時，通常會新增幾個節點，測試該點的串流，可能新增節點以移出部分表狀或圖形輸出，然後繼續建置串流。 通常，您無法承擔每次測試串流時都執行海量資料的資料傳遞。

建立海量資料的適用資料範本，可讓您針對實際資料測試串流，而不會導致在執行完成資料傳遞時需要的時間延遲。 資料範例必須包含足夠的資料，才能順利執行串流。 例如，如果您計劃分析位於明尼蘇達州之儲存庫的交易，則資料範例必須包含位於明尼蘇達州之儲存庫的交易。

取樣之後，您可以：

- 建立海量資料所在之叢集上資料範例的快取，或者

  Pros - 簡式且不需要切換來源節點

  Cons - 階段作業結束後快取消失
- 建立包含資料範例的新 Analytic Server 資料來源，或者

  Pros - 永久資料來源

  Cons - 需要編輯/切換來源節點
- 將資料範例下載至本端系統，並建立本端資料來源

  Pros - 進行原型設計時不耗用叢集資源；當您使用少量資料時，SPSS Modeler 用戶端比 Analytic Server 更有效。

  Cons - 需要切換來源節點

從「來源」節點建立個別「類型」與「過濾器」節點

每個 SPSS Modeler 來源節點還具有結合的「過濾器」與「類型」節點的功能。 這有助於保持畫布簡化，但是在您切換至不同的「來源」節點類型時會較難。 此外，它會遮蔽正在發生「類型」與「過濾器」作業這一事實。

將「過濾器」與「選取」節點盡量放置在接近「來源」節點的位置

這會減少下游作業中的記錄數目。

盡可能避免「排序」節點

根據所排序的資料，Analytic Server 不支援節點中的最佳化（例如「合併」節點）。 因此，中游「排序」節點很少有用。 當「排序」節點後面緊接著「範例」節點時，會具有值，以取得「前 N 筆」（或者「後 N 筆」）記錄。

僅計算將使用的欄位

不計算某個欄位，然後立即對它進行過濾。

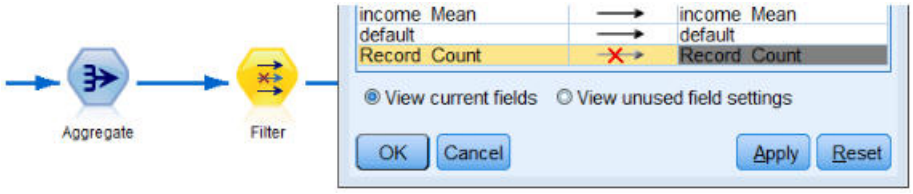

圖 *6. Modeler* 欄位選項

只要可能，盡量不設計難以理解的表示式，避免建立大量暫時欄位。 例如，不定義下列範例：

```
now = datetime_now()
birthdate = datetime_date(bYear, bMonth, bDay)
age = date_years_difference(birthdate, now)
```

而是定義下列範例：

```
age = date_years_difference(datetime_date(bYear, bMonth, bDay), datetime_now())
```

以此方法將暫存項摺疊至行內表示式，可能會在轉換大量欄位時提高效能。

在資料來源中設定儲存體

變更欄位的儲存體類型（例如，字串至整合）中游的作業可能有害於整體效能。 在「Analytic Server 主控台」中定義資料來源時，您可以設定欄位的儲存體，以避免重複這些轉換。

當您使用少量資料時，使用 SPSS Modeler

使用 Analytic Server 操作海量資料，然後使用 SPSS Modeler 以完成少量資料的計算。

選取適當的 Analytic Server 相關串流內容

配置相關的串流內容（**工具** > **選項** > **串流內容** > **分析伺服器**），決定是否容許資料處理退出 Analytic Server 並在 SPSS Modeler 中繼續（節點無法在 Analytic Server 中執行）。

依預設，SPSS Modeler 配置為報告錯誤，並停止在此狀況下執行： 您可以將設定從錯誤變更為警告，並調整可以在 SPSS Modeler 中處理的資料量限制，從而略過該錯誤。 例如，您可以從預設 10000 記錄值更新資料傳送速率（必要的話）。 請注意，檢視使用 SPSS Modeler 表格節點的結果時，此限制也適用。 如果超出該限制，則「資料」提取的 SPSS Modeler 報告已超出串流內容中設定的限制。

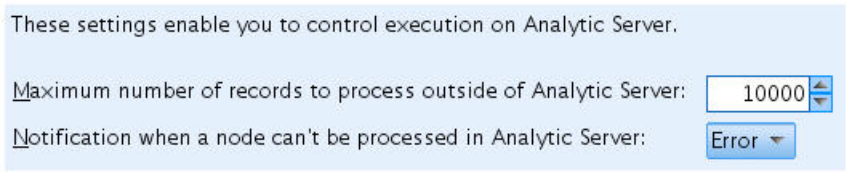

圖 *7. Analytic Server* 設定

使用 Analytic Server 來源節點

Analytic Server 可以連接至不同的資料庫資料來源，但是 SPSS Modeler 需要所有「來源」節點都是 Analytic Server 的「來源」節點（從而讓整個串流作為 Analytic Server 工作執行）。對於在 Analytic Server 中執行的整個串流，資料來源節點必須變更為 Analytic Server 的「來源」節點，並且必須在「Analytic Server 主控台」中建立 Analytic Server 資料庫資料來源。

考量如何使用受支援的節點

Analytic Server 不支援所有節點（「移轉」節點是良好的範例）。 為了合併移轉作業的結果與剩餘串流，並讓它在 Analytic Server 中執行，包括「移轉」節點的子串流應該寫出使用 Analytic Server 之「匯出」節點的 Analytic Server 資料來源。 然後，您可以連接 Analytic Server 的「來源」節點，其中岔斷串流以寫入 Analytic Server。

註：移轉作業適合一次性或很少執行的作業，但是應該用於常式串流作業。

判定串流是否先在 Analytic Server 中工作，然後再執行

準備串流以在 Analytic Server 中執行之後，選取終端機節點並使用 SPSS Modeler 預覽特性（工具列上的**預覽執行**控制項），以驗證執行終端機節點涉及之任何節點是否將在 Analytic Server 中運作（而不執行串流）。系統會在訊息視窗中報告問題。

結合背對背合併作業

當一系列「合併」節點具有相同的金鑰及結合類型時，可以將它們與單一節點結合。

結合相同的子串流

盡可能嘗試合併相同的子串流，尤其是當它們包含昂貴的作業時（例如，合併及排序）。 SPSS Modeler 會一次性執行這些作業，並使用快取以改進效能。 在下列範例中，根據 **newField** 節點，串流相同。

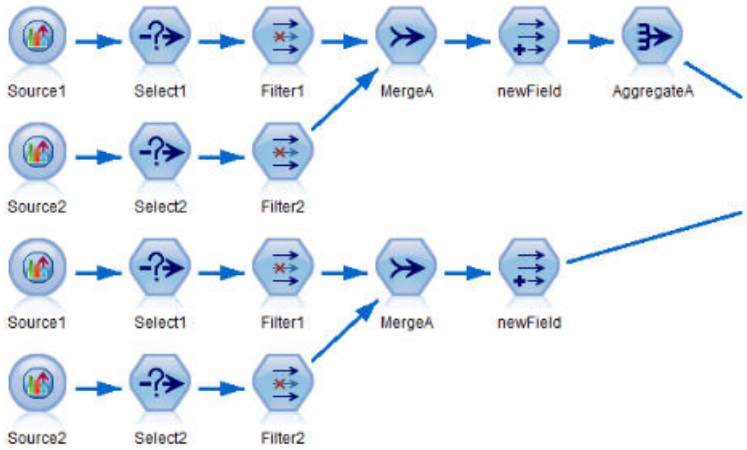

圖 *8.* 範例串流

如果子串流改為如下所示結構化，則它會更有效（且更易於維護）：

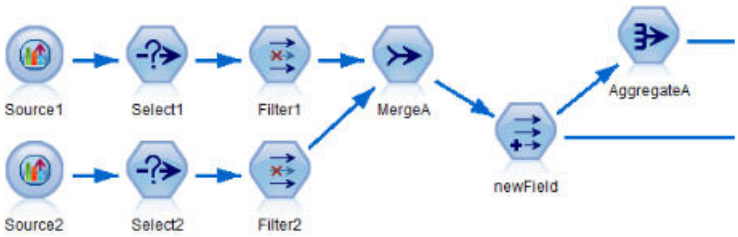

圖 *9.* 範例串流

移除額外的類型節點

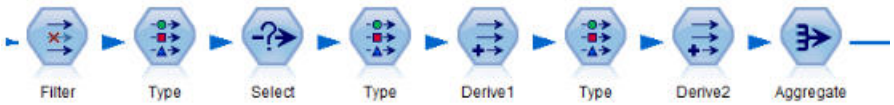

圖 *10.* 範例串流

在針對 Analytic Server 執行時，避免不必要的「類型」節點。 「類型」節點的讀取值作業會啟動 MapReduce 工作。 這通常是一次性節省項，除非您清除「類型」節點值。

完整記錄每一個串流

下列範例顯示包含許多子串流的複式串流。

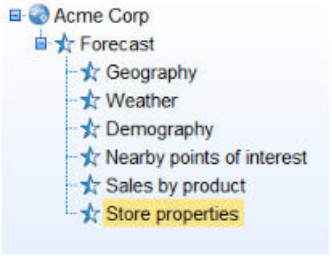

圖 *11.* 子串流範例

在這種情況下，請務必適當地命名超級節點並記錄串流（正如您記錄程式碼一樣）。 清除註解可能為讀取或維護串流的其他分析師提供無價值的資訊。 例如：

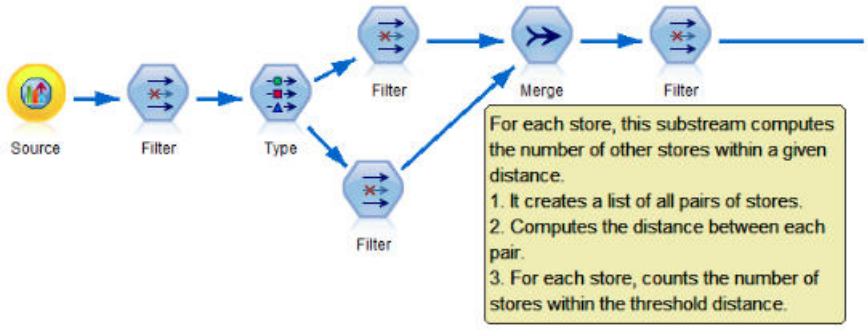

圖 *12.* 具有註解的串流範例

開發串流時，使用 SPSS Modeler 快取以快速地儲存中間結果

在針對 Analytic Server 執行的串流中，透過將串流之特定部分中的資料儲存至 HDFS 上的暫存檔（與儲存在 SPSS Modeler 伺服器上完全不同），節點快取運作。 快取非常適合海量資料，且可安全地使用在 Analytic Server 上執行的串流中。

第 5 章 疑難排解

Analytic Server 提供數個有用的工具來協助判斷問題。

記載

Analytic Server 會建立客戶日誌檔和追蹤檔案，對問題診斷會有幫助。 藉由預設的 Liberty 安裝，您可以在 {AS_ROOT}/ae_wlpserver/usr/servers/aeserver/logs 目錄中找到日誌檔。

預設記載配置會產生兩個日誌檔（每天輪換）。

as.log
此檔案包含參考資訊警告及錯誤訊息的高層次摘要。 當無法利用使用者介面中所顯示錯誤訊息來解決發生的伺服器錯誤時，應該首先檢查該檔案。

as_trace.log
此檔案包含 ae.log 中的所有項目，但會新增主要針對 IBM 支援中心及用於開發除錯的其他資訊。

Analytic Server 使用 Apache LOG4J 作為其基礎記載機能。 藉由 LOG4J，記載可以透過編輯 {AS_SERVER_ROOT}/configuration/log4j.xml 配置檔進行動態調整。 支援中心可能會要求您執行此動作來協助論斷問題，或要求您修改它來限制保留的日誌檔數目。 系統會在數秒內自動偵測檔案變更，因此 Analytic Server 不需要重新啟動。

如需 log4j 及配置檔的相關資訊，請參閱官方 Apache 網站 http://logging.apache.org/log4j/ 的說明文件。

版本資訊

您可以檢查 {AS_ROOT}/properties/version 資料夾，來確定安裝哪個版本的 Analytic Server。 下列檔案包含版本資訊。

IBM_SPSS_Analytic_Server-*.swtag
包含詳細產品資訊。

version.txt
已安裝產品的版本和建置號碼。

日誌收集器

當無法直接檢閱日誌檔來解決問題時，您可以組合所有日誌並將其傳送至 IBM 支援中心。 提供了公用程式，以便輕鬆收集所有必要的資料。

使用指令 Shell，來執行下列指令：

```
cd  {AS_ROOT}/bin
run  >sh ./logcollector.sh
```

這些指令會在 {AS_ROOT}/bin 下建立壓縮檔。 壓縮檔包含所有日誌檔及產品版本資訊。

常見問題

本節說明了部分常見管理問題，以及如何對其進行修正。

執行中串流

R 工作將非英文單字轉換為 Unicode

在 Cloudera 叢集上，如果 Hadoop 伺服器的系統編碼不是 UTF-8，則 R 會將非英文單字轉換為 Unicode。

1. 導覽至 Cloudera Manager 主控台中的 YARN 配置標籤。
2. 將下列設定新增至「NodeManager 環境進階配置 Snippet（安全閥）」欄位。

```
LC_ALL=""
LANG=en_US.utf8
```

PySpark 工作無法執行

確保 Spark 服務部署在所有 Analytic Server 節點及所有節點管理程式中。

PySpark 工作無法在啟用了 Kerberos 的環境上執行

您必須執行 `kinit` 指令，並重新啟動 Analytic Server，然後 PySpark 測試才能順利執行。 例如：

HDP Kerberos

```
cd /etc/security/keytabs/
sudo -u as_user kinit -k -t as_user.headless.keytab as_user/lyrh1.fyre.ibm.com@IBM.COM
```

CDH Kerberos

```
cd /run/cloudera-scm-agent/process/387-analyticserver-ANALYTIC_SERVER
sudo -u as_user kinit -k -t analyticserver.keytab as_user/cdh12-1.fyre.ibm.com@IBM.COM
```

記憶體發生錯誤

在執行程式記憶體發生錯誤之後配置 YARN

當所需執行程式記憶體大於臨界值上限時，可能發生下列錯誤：

```
Caused by: com.spss.mapreduce.exceptions.JobException:
    java.lang.IllegalArgumentException: Required executor memory (1024+384  MB) is above the max
    threshold (1024 MB) of this cluster! Please  increase the value of
    'yarn.scheduler.maximum-allocation-mb'.
```

下列步驟提供了為解決這個問題所需採用的 YARN 配置設定。

針對 Ambari

1. 在 Ambari 使用者介面中，跳至 **YARN** > **配置** > **設定**。
2. 將記憶體**節點（為所有 YARN 儲存器配置的記憶體）**增加到 8192MB。
3. 增加儲存器值：
   - **儲存器大小下限（記憶體）**增加到 682MB
   - **儲存器大小上限（記憶體）**增加到 8192MB
4. 將**儲存器大小上限 (VCore)** 增加到 3。
5. 重新啟動 YARN、Spark 及 Analytic Server 服務。

針對 Cloudera

1. 將 `yarn.nodemanager.resource.memory-mb` 增加到 8GB
   - 在 Cloudera Manager 使用者介面中，跳至**YARN 服務** > **配置** > **搜尋儲存器記憶體**，然後將值增加到 8GB。
2. 在 Cloudera Manager 使用者介面中，跳至**YARN 服務** > **快速鏈結**，然後選取**動態資源儲存區**。
3. 在**配置**下，針對每一個可用的儲存區按一下**編輯**，然後在 **YARN** 下，將**執行中應用程式上限**值設為 4。
4. 重新啟動 YARN、Spark 及 Analytic Server 服務。

Hadoop 與 Apache Spark 2.x

- 當同一環境中存在 Hadoop 與 Apache Spark 2.x 時，大部分 forcespark 及 forcehadoop 工作會失敗。該錯誤在 Yarn 應用程式日誌中顯示為： java.lang.NoClassDefFoundError: org/apache/hadoop/fs/FSDataInputStream。

  依如下所示手動編輯 /etc/spark2/conf/spark-defaults.conf 檔案可解決該問題：

  ```
  #spark.hadoop.mapreduce.application.classpath=
  #spark.hadoop.yarn.application.classpath=
  ```

- 當同一系統上安裝了兩個 JDK 版本時，Cloudera 會使用 JDK 1.7，而 Spark 2.x 會使用 JDK 1.8。 使用 Apache Spark 2.x 執行 forcespark 或 forcehadoop 工作可能導致所有工作失敗，並產生下列錯誤訊息：

  執行失敗。 原因：org/apache/spark/api/java/function/PairFunction：major.minor 52.0 版不受支援的

  對於 Cloudera，請在 Cloudera Manager 之 **server.env 的 Analytic Server 進階配置 Snippet（安全閥）**區段中新增下列行：

  ```
  JAVA_HOME=/usr/java/jdk1.8.0_152
  ```

將 *admin* 權限授與 Apache Hive UDF 使用者

登錄 Analytic Server Apache Hive UDF 之後，您可能遇到功能無效錯誤。依預設，有兩個 Hive 角色（admin 與 public）。Hive 使用者屬於 public 角色。Hive UDF 需要已登錄使用者具有 admin 專用權（已啟用 Hive 安全）。

若要將 admin 權限授與 Hive UDF 使用者：

1. 以 Hive 身份登入 Beeline：

   ```
   !connect jdbc:hive2://localhost:10000/default;principal=hive/cdh51501.fyre.ibm.com@IBM.COM
   ```

2. 在 Beeline 中執行下列指令：

   ```
   grant admin to user hive WITH ADMIN OPTION;
   ```

   註：其他有用的 SQL 指令包括：

   顯示哪些角色已指派給使用者 hive

   ```
   show role grant user hive;
   ```

   顯示哪些使用者已指派給 public 角色

   ```
   show principals public;
   ```

3. 重新啟動 Hive 並重新登錄 Analytic Server Hive UDF。

   ```
   sudo -u hive kinit -k -t hive.keytab hive/cdh51501.fyre.ibm.com@IBM.COM
   sudo -u hive hive -f /opt/cloudera/parcels/AnalyticServer/bin/udfUnregister.sql
   sudo -u hive hive -f /opt/cloudera/parcels/AnalyticServer/bin/udfRegister.sql
   ```

HiveDB 錯誤

寫入至 HiveDB 時，您可能遇到下列錯誤：

```
(AEQAE4805E) 執行失敗。原因：com.google.common.io.Closeables.closeQuietly(Ljava/io/Closeable;)
```

該錯誤是由 Hadoop Cluster 上多個版本的 guava-*.jar 檔造成的。可以透過執行下列步驟解決該錯誤（該範例使用 HDP 3.1）：

1. 開啟 Ambari 主控台並停止 Analytic Server 服務。

2. 將 /usr/hdp/3.1.0.0-78/spark2/jars/guava-14.0.1.jar 複製到 {AS_ROOT}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib。
3. 在 Ambari 主控台中，重新整理 Analytic Server 服務，然後啟動 Analytic Server 服務。

效能調整

本節說明最佳化系統效能的方法。

Analytic Server 是 Ambari 架構中的一個元件，該元件會利用其他元件，例如 HDFS、Yarn 及 Spark。Hadoop、HDFS 及 Spark 的一般效能調整技術適用於 Analytic Server 工作量。 每個 Analytic Server 工作量都有所不同，因此，需要根據您的特定部署工作量進行調整實驗。 下列內容及調整提示是主要變更，已影響 Analytic Server 評比及調整大小測試的結果。

當第一個工作在 Analytic Server 上執行時，伺服器將啟動持續性 Spark 應用程式，該應用程式將處於作用中，直到 Analytic Server 關閉為止。 在 Analytic Server 執行期間，持續性 Spark 應用程式將在所有配置給它的叢集資源上進行配置並予以保留，即使 Analytic Server 工作未在積極地執行中。 應該小心計算配置給 Analytic Server Spark 應用程式的資源量。 如果所有叢集資源都配置給 Analytic Server Spark 應用程式，則其他工作可能會延遲或未執行。 這些工作可以排入佇列中等待足夠的可用資源，並且 Analytic Server Spark 應用程式將耗用那些資源。

如果配置並部署多個 Analytic Server 服務，則每一個服務實例都可以潛在配置其自己的持續性 Spark 應用程式。 例如，如果部署兩個 Analytic Server 服務以支援高可用性失效接手，則您可以看到兩個持續性 Spark 應用程式在作用中，每一個都配置叢集資源。

其他複雜性是在某些狀況下，Analytic Server 可能啟動需要叢集資源的對映減少工作。 這些對映減少工作將需要未配置給 Spark 應用程式的資源。 需要對映減少工作的特定元件是 PSM 模型建置。

可以配置下列內容，以將資源配置給 Spark 應用程式。 如果在 Spark 安裝的 spark-defaults.conf 中設定它們，則可以為環境中執行的所有 Spark 工作配置它們。 如果在 Analytic Server 配置中將它們設為「自訂 analytic.cfg」區段下的自訂內容，則只會為 Analytic Server Spark 應用程式配置它們。

spark.executor.memory

每個執行程式處理程序使用的記憶體數量。

spark.executor.instances

要啟動的執行程式處理程序數目。

spark.executor.cores

每個執行程式處理程序的執行程式工作程式執行緒數目。 此值應該介於 1 和 5 之間。

設定三個主要 Spark 內容的範例。 HDFS 叢集中有 10 個資料節點，每一個資料節點有 24 個邏輯核心和 48 GB 記憶體，並且僅執行 HDFS 處理程序。 這裡有一個配置此環境內容的方法，假設您僅在此環境上執行 Analytic Server 工作，並且想要對單一 Analytic Server Spark 應用程式配置上限。

- 設定 spark.executor.instances=20。 這將嘗試在每個資料節點上執行 2 個 Spark 執行程式處理程序。
- 設定 spark.executor.memory=22G。 這會將每一個 Spark 執行程式處理程序的資料堆大小上限設為 22 GB，並在每一個資料節點上配置 44 GB。 其他 JVM 及作業系統需要額外記憶體。
- 設定 spark.executor.cores=5。 這將為每一個 Spark 執行程式提供 5 個工作程式執行緒，並為每個資料節點提供總計 10 個工作程式執行緒。

監視用來執行工作的 Spark 使用者介面

如果您看到可能影響效能的磁碟「溢出」。 部分可能的解決方案為：

- 透過 **spark.executor.memory** 提高記憶體並將它配置給 Spark 執行程式。
- 減少 **spark.executor.cores** 的數目。 這將減少配置記憶體的並行工作執行緒的數目，但是它還會減少工作的平行化數量。
- 變更 Spark 記憶體內容。 Spark 之 Spark 執行程式資料堆的 **spark.shuffle.memoryFraction** 與 **spark.storage.memoryFraction** 配置百分比。

確保相同的節點具有足夠的記憶體

如果 HDFS 中的區塊數目較大並在不斷增長，請確保您指定的節點資料堆增加以容納此增長。 這是一般 HDFS 調整建議。

變更用於快取的記憶體數量

依預設，**spark.storage.memoryFraction** 的值為 0.6。 當資料的 HDFS 區塊大小為 64MB 時，此可以增加達 0.8。 如果輸入資料的 HDFS 區塊大小大於 64MB，則僅當每個作業配置的記憶體大於 2GB 時，才可以增加此值。

調整模型評分的效能

您可以透過使用下列步驟，在具有 Apache Spark 引擎的海量資料集上，改進模型評分工作的效能。 請注意，這些步驟不得影響叢集上非 Analytic Server 服務的作業。

1. 檢查 libtcmalloc_minimal.so{/version} 是否已安裝在叢集中的每一個節點上。

   ```
   whereis libtcmalloc_minimal.so.*
   ```

2. 如果未安裝 libtcmalloc_minimal.so，請在叢集中的每一個節點上安裝包含 libtcmalloc_minimal 程式庫的作業系統特定套件，或者手動建置並安裝 libtcmalloc_minimal。 例如：

   Ubuntu：

   ```
   sudo apt-get install libgoogle-perftools-dev
   ```

   Red Hat Enterprise Linux 6.x (x64)：

   a. 為 RedHat 安裝 EPEL 儲存庫（如果尚未安裝）

      ```
      wget http://dl.fedoraproject.org/pub/epel/6/x86_64/epel-release-6-8.noarch.rpm
      sudo rpm -Uvh epel-release-6*.rpm
      ```

   b. `sudo yum install gperftools-libs.x86_64`

   手動建置：

   a. 從鏈結 https://github.com/gperftools/gperftools/releases 中下載 gperftools-2.4.tar.gz

   b. `tar zxvf gperftools-2.4.tar.gz`

   c. `cd gperftools-2.4`

   d. `./configure --disable-cpu-profiler --disable-heap-profiler --disable-heap-checker --disable-debugalloc --enable-minimal`

   e. `make`

   f. `sudo make install`

3. 從一個以上節點上執行的下列指令中返回時，請記錄所安裝程式庫檔案 libtcmalloc_minimal.so{.version} 的其中一個位置。

```
whereis libtcmalloc_minimal.so.*
```

如果叢集有節點正在執行混合作業系統，則此檔案可能有多個位置。

4. 在 Ambari 主控台中，前往 Analytic Server 配置，並在「自訂 analytics.cfg」區段下，將程式庫的位置用作值，來配置主要 spark.executorEnv.LD_PRELOAD。 進行此變更後，重新啟動 Analytic Server 服務。 例如，如果程式庫安裝至 /usr/lib64/libtcmalloc_minimal.so.4，則配置將是：

   ```
   spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4
   ```

   如果需要多個位置，請使用空格來區隔它們，如下列範例中所示。

   ```
   spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4 /usr/lib/libtcmalloc_minimal.so
   ```

   如果有任何節點未在其中一個所配置的位置上安裝 libtcmalloc_minimal.so 程式庫，這不會造成錯誤，但是這些節點上模型評分的效能可能更為緩慢。

Spark 對映端結合

Analytic Server Spark 結合實作不支援對映端結合功能（Spark 結合主要在減少端）。 實作不會利用對映端結合以在一個輸入較小時最佳化結合。 不利用對映端結合會導致最終失敗的極端資源密集型程式 Spark 工作。

若要在執行 Analytic Server Spark 對映端結合（或者基於最小 RDD 大小的原生 Spark 工作）時最佳化結合，您可以將 spark.msj.maxBroadcast 內容新增至 analytics.cfg 檔 (SPSS Analytic Server/Configs/Custom analytics.cfg) 或 analytics-meta。

注意事項

本資訊係針對 IBM 在美國所提供之產品與服務所開發。 IBM 可能以其他語言提供本資訊。 不過，您可能需要擁有一份該語言的產品或產品版本的副本，才能存取該產品或產品版本。

在其他國家，IBM 不見得有提供本文件所提及之各項產品、服務或功能。 請洽詢當地的 IBM 業務代表，以取得當地目前提供的產品和服務之相關資訊。 本文件在提及 IBM 的產品、程式或服務時，不表示或暗示只能使用 IBM 的產品、程式或服務。 只要未侵犯 IBM 之智慧財產權，任何功能相當之產品、程式或服務皆可取代 IBM 之產品、程式或服務。 不過，任何非 IBM 之產品、程式或服務，使用者必須自行負責作業之評估和驗證責任。

本文件所說明之主題內容，IBM 可能擁有其專利或專利申請案。 提供本文件不代表提供這些專利的授權。 您可以書面提出授權查詢，來函請寄到：

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

如果是有關雙位元組 (DBCS) 資訊的授權查詢，請洽詢所在國的 IBM 智慧財產部門，或書面提出授權查詢，來函請寄到：

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

International Business Machines Corporation 只依「現況」提供本出版品，不提供任何明示或默示之保證，其中包括且不限於不侵權、可商用性或特定目的之適用性的隱含保證。 有些地區在特定交易上，不允許排除明示或暗示的保證，因此，這項聲明不一定適合您。

本資訊中可能會有技術上或排版印刷上的訛誤。因此，IBM 會定期修訂；並將修訂後的內容納入新版中。 IBM 隨時會改進及/或變更本出版品所提及的產品及/或程式，不另行通知。

本資訊中任何對非 IBM 網站的敘述僅供參考，IBM 對該網站並不提供任何保證。 這些網站所提供的資料不是 IBM 本產品的資料內容，如果要使用這些網站的資料，您必須自行承擔風險。

IBM 得以各種 IBM 認為適當的方式使用或散布 貴客戶提供的任何資訊，而無需對 貴客戶負責。

如果本程式之獲授權人為了 (i) 在個別建立的程式和其他程式（包括本程式）之間交換資訊，以及 (ii) 相互使用所交換的資訊，因而需要相關的資訊，請洽詢：

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
US

上述資料之取得有其特殊要件，在某些情況下必須付費方得使用。

IBM 基於 IBM 客戶合約、IBM 國際程式授權合約或雙方之任何同等合約的條款，提供本文件所提及的授權程式與其所有適用的授權資料。

所述的效能資料及客戶範例僅供示範之用。 實際的效能結果可能會因特定的配置及作業狀況而異。

本文件所提及之非 IBM 產品資訊，取自產品的供應商，或其發佈的聲明或其他公開管道。 IBM 並未測試過這些產品，也無法確認這些非 IBM 產品的執行效能、相容性或任何對產品的其他主張是否完全無誤。 有關非 IBM 產品的性能問題應直接洽詢該產品供應商。

所有關於 IBM 未來方針或目的之聲明，隨時可能更改或撤銷，不必另行通知，且僅代表目標與主旨。

所有 IBM 價格都是 IBM 建議的零售價格，可隨時變更而不另行通知。 經銷商價格可不同。

本資訊僅作規劃目的。 在產品可用前，此處的資訊可能變更。

本資訊含有日常商業運作所用之資料和報告範例。 為了盡可能地加以完整說明，範例中含有個人、公司、品牌及產品的名稱。 此等名稱皆屬虛構，凡有類似實際個人或企業所用之名稱及地址者，皆屬巧合。

著作權：

本資訊含有日常商業運作所用之資料和報告範例。 為了盡可能地加以完整說明，範例中含有個人、公司、品牌及產品的名稱。 此等名稱皆屬虛構，凡有類似實際個人或企業所用之名稱及地址者，皆屬巧合。

這些範例程式或任何衍生著作的每份副本或任何部分，都必須依照下列方式併入著作權聲明：

商標

IBM、IBM 標誌及 ibm.com 是 International Business Machines Corp. 在世界許多管轄區註冊的商標或註冊商標。 其他產品及服務名稱可能是 IBM 或其他公司的商標。 IBM 商標的最新清單可在 Web 的 "Copyright and trademark information" 中找到，網址為 www.ibm.com/legal/copytrade.shtml。

Adobe、Adobe 標誌、PostScript 及 PostScript 標誌是 Adobe Systems Incorporated 在美國及（或）其他國家或地區的註冊商標或商標。

IT Infrastructure Library 是 Central Computer and Telecommunications Agency（現在是 Office of Government Commerce 的一部分）的註冊商標。

Intel、Intel 標誌、Intel Inside、Intel Inside 標誌、Intel Centrino、Intel Centrino 標誌、Celeron、Intel Xeon、Intel SpeedStep、Itanium 及 Pentium 是 Intel Corporation 或其子公司在美國及其他國家或地區的商標或註冊商標。

Linux 是 Linus Torvalds 在美國及（或）其他國家或地區的註冊商標。

Microsoft、Windows、Windows NT 及 Windows 標誌是 Microsoft Corporation 在美國及/或其他國家或地區的商標。

ITIL 是 Minister for the Cabinet Office 在美國 Patent and Trademark Office 註冊的註冊商標及註冊社群商標。

UNIX 是 The Open Group 在美國及其他國家或地區的註冊商標。

Cell Broadband Engine 是 Sony Computer Entertainment, Inc. 在美國及/或其他國家或地區的商標並在當地軟體使用權下使用。

Linear Tape-Open、LTO、LTO 標誌、Ultrium 及 Ultrium 標誌是 HP、IBM Corp. 及 Quantum 在美國及其他國家的商標。

Printed in Taiwan