IBM SPSS Analytic Server
Versão 3.2.2

Guia do Administrador

Nota

Antes de utilizar estas informações e o produto suportado por elas, leia as informações em “Avisos” na página 39.

Informações sobre o Produto

Esta edição se aplica à versão do 3, liberação 2, modificação 2 do IBM® SPSS Analytic Server e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. Gerenciamento de Arrendatário

Os locatários fornecem uma divisão de alto nível de usuários, projetos e origens de dados para que os objetos não possam ser compartilhados entre locatários. Cada usuário acessa o sistema no contexto de um arrendatário para o qual ele é designado.

Você gerencia locatários e designa usuários a locatários, no console do Analytic Server. A visualização da página Locatários depende da função do usuário que está registrada no console:

- O administrador "super usuário" que é configurado durante a instalação é o gerenciador de locatários. Apenas esse usuário pode criar novos arrendatários e editar as propriedades de qualquer arrendatário.
- Os usuários com função de Administrador podem editar as propriedades do arrendatário ao qual estão conectados.
- Os usuários com função de Usuário não podem editar propriedades do arrendatário. A página Locatários é ocultada deles.
- Usuários com a função de Leitor não podem editar origens de dados ou mesmo efetuar login no console do Analytic Server.

Os administradores podem acessar as páginas Projetos e Origens de Dados e gerenciar qualquer projeto ou origem de dados para limpeza e administração. Consulte o *IBM SPSS Analytic Server Guia do Usuário* para obter mais informações.

Lista de locatários

A página principal Locatários exibe os locatários existentes em uma tabela. Apenas o administrador "super usuário" pode fazer edições nessa página.

- Clique no nome de um locatário para exibir seus detalhes e editar suas propriedades.
- Clique na URL de um locatário para abrir o console no contexto desse locatário.

  Nota: Você será desconectado do console e precisará efetuar login com credenciais válidas para o locatário.
- Digite na área de procura para filtrar a lista a fim de exibir apenas os locatários com a sequência de caracteres de procura em seu nome.
- Clique em **Novo** para criar um novo locatário com o nome especificado no diálogo **Incluir novo locatário**. Consulte “Regras de nomenclatura ” na página 2 para restrições nos nomes que você pode atribuir aos locatários.
- Clique em **Excluir** para remover o(s) locatário(s) selecionado(s).
- Clique em **Atualizar** para atualizar a lista.

Detalhes de locatário individual

A área de conteúdo é dividida em várias seções reduzíveis.

Detalhes

Nome
Um campo de texto editável que exibe o nome do arrendatário.

Descrição
Um campo de texto editável que permite fornecer texto explicativo sobre o arrendatário.

URL
Esta é a URL a ser fornecida aos usuários para efetuarem login no arrendatário através do console do Analytic Server, e a ser usada para configurar o servidor SPSS Modeler. Consulte *IBM SPSS Analytic Server Installation and Configuration Guide* para obter detalhes sobre configurar o SPSS Modeler.

Status

Ativo os locatários estão atualmente em uso. Tornar um locatário **Inativo** evita que os usuários efetuem login para esse locatário, mas não exclui nenhuma informação subjacente.

Diretores

Os diretores são usuários e grupos extraídos do provedor de segurança que é configurado durante a instalação. É possível incluir diretores em um locatário como Administradores, Usuários ou Leitores.

- Digitando os filtros da caixa de texto nos usuários e grupos com a sequência de caracteres de procura em seu nome. Selecione **Administrador**, **Usuário** ou **Leitor** na lista suspensa para designar suas funções dentro do locatário. Clique em **Incluir participante** para incluí-lo na lista de autores.
- Para remover um participante, selecione um usuário ou grupo na lista de membros e clique em **Remover participante**.

Métricas

Permite configurar limites de recurso para um locatário. Relata o espaço em disco atualmente usado pelo locatário.

- É possível configurar uma cota de espaço em disco máxima para o locatário; quando este limite é atingido, mais nenhum dado pode ser gravado no disco neste locatário até que seja liberado espaço em disco suficiente para colocar o uso do espaço em disco do locatário abaixo da cota.
- É possível configurar um nível de aviso de espaço em disco para o locatário; quando a cota é excedida, nenhuma tarefa analítica pode ser enviada por diretores neste locatário até que seja liberado espaço em disco suficiente para colocar o uso do espaço em disco do locatário abaixo da cota.
- É possível configurar um número máximo de tarefas paralelas que podem ser executadas uma única vez neste locatário; quando a cota é excedida, nenhuma tarefa analítica pode ser enviada por diretores neste locatário até que uma tarefa em execução atualmente seja concluída.
- É possível configurar o número máximo de campos que uma origem de dados pode ter. O limite é verificado sempre que uma origem de dados é criada ou atualizada.
- É possível configurar o tamanho máximo do arquivo em megabytes. O limite é verificado quando um arquivo é atualizado.

Configuração do provedor de segurança

Permite especificar o provedor de autenticação do usuário. **Padrão** usa o provedor do locatário padrão, que foi configurado durante a instalação e a configuração. **LDAP** permite autenticar usuários com um servidor LDAP externo, como Active Directory ou OpenLDAP. Especifique as configurações para o provedor e opcionalmente as configurações de filtro para controlar os usuários e grupos disponíveis na seção Diretores.

Regras de nomenclatura

Para qualquer coisa que possa receber um nome exclusivo no Analytic Server, como origens de dados e projetos, as regras a seguir são aplicadas a esses nomes.

- Em um único locatário, os nomes devem ser exclusivos dentro de objetos do mesmo tipo. Por exemplo, duas origens de dados não podem ser denominadas insuranceClaims, mas uma origem de dados e um projeto poderiam cada um ser denominados insuranceClaims.
- Os nomes fazem distinção entre maiúsculas e minúsculas. Por exemplo, insuranceClaims e InsuranceClaims são considerados nomes exclusivos.
- Os nomes ignoram espaço em branco à esquerda e à direita.
- Os caracteres a seguir são inválidos nos nomes.

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

Capítulo 2. Introdução para usuários

Informe aos usuários para navegar até `http://<host>:<port>/<context-root>/admin/<tenant>` e inserir o nome de usuário e a senha para efetuar logon no console do Analytic Server.

Nota: O nome de usuário inserido durante o prompt de login do console do Analytic Server é inserido sem o sufixo do nome da região. Como resultado, quando diversas regiões são definidas, os usuários visualizam uma lista suspensa de **Regiões** que os permite selecionar a região adequada. Quando somente uma região é definida, os usuários não visualizam uma lista suspensa de **Regiões** ao conectar-se ao Analytic Server.

<host>
O endereço do host do Analytic Server.

<porta>
A porta pela que Analytic Server está recebendo. Por padrão, este é 9080.

<context-root>
A raiz de contexto do Analytic Server. Por padrão, é um analyticserver.

<tenant>
Em um ambiente de diversos locatários, o locatário ao qual você pertence. Em um ambiente de único locatário, o locatário padrão é **`ibm`**.

Por exemplo, se a máquina host tem endereço IP 9.86.44.232, você criou um locatário "mycompany" e incluiu usuários nele e as outras configurações foram deixadas com os seus padrões; então, os usuários devem navegar para `http://9.86.44.232:9080/analyticserver/admin/mycompany` para acessar o console do Analytic Server.

Capítulo 3. Nomes de tarefa do Analytic Server

O Analytic Server produz tarefas de redução de mapa e Spark, que podem ser monitoradas por meio da sua interface com o usuário de Gerenciador de Recursos do cluster Hadoop.

O nome da tarefa de redução de mapa possui a estrutura a seguir.

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

{tenant name}
Este é o nome do locatário sob o qual a tarefa é executada.

{user name}
Este é o usuário que solicitou a tarefa.

{algorithm name}
Este é o algoritmo primário na tarefa. Observe que um único fluxo pode gerar diversas tarefas de redução de mapa; da mesma forma, várias operações em um fluxo podem estar contidas em uma única tarefa de redução de mapa.

Todas as tarefas de redução de mapa são exibidas na interface com o usuário do Gerenciador de Recursos. Um aplicativo Spark único é iniciado para cada Analytic Server. Abra a interface com o usuário do aplicativo Spark para monitorar tarefas Spark (os nomes de Tarefa são exibidos na coluna **Descrição**).

Capítulo 4. Propriedades Customizadas do Analytic Server

As propriedades customizadas a seguir são definidas ou podem ser configuradas no arquivo `analytic.cfg` do Analytic Server.

Tabela 1. Propriedades customizadas do Analytic Server

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
admin.username	Sequência			Define o nome do usuário administrativo do Analytic Server.
ae.cluster.ha.cascade.failure.protection	Booleano	VERDADEIRO		Quando ativado (true), o gerenciador de trabalho do cluster evita que as tarefas que falharam em vários membros de cluster travem todos os servidores de cluster. Isso é feito suspendendo permanentemente a tarefa incorreta ou atribuindo-a para ser executada apenas em um servidor de quarentena designado.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
ae.cluster.heapdump_onexit.filename	Sequência			Quando um nome do arquivo é especificado, a JVM do Analytic Server grava um dump do heap (para o arquivo especificado) quando vários eventos **CPU_Starvation** ocorrem.
ae.cluster.job.artifacts.cleanup.delay.minutes	Número inteiro	5		A quantidade de tempo que o Analytic Server espera após a conclusão de uma tarefa antes de limpar os artefatos relacionados à tarefa no zookeeper.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
ae.cluster.quarantine.server.name	Sequência			Em ambientes em cluster, identifica o servidor do Analytic Server que é usado para executar tarefas em quarentena (tarefas que excedem o limite de contagem de falhas).
ae.cluster.queue.callback.threadpool.size	Número inteiro	20		Tamanho do **ThreadPool** que é usado pelo serviço de cluster ao consumir mensagens de cluster entre processos enviadas por meio do zookeeper.
ae.cluster.thread.scheduler.delay.detector	Número inteiro	30		Detecta problemas de desempenho local (por exemplo, registro de mensagens de falta de CPU).

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
ae.cluster.thread.scheduler.detect.error	Número inteiro	10		Detecta problemas de desempenho local (por exemplo, registro de mensagens de falta de CPU).
as.db.connect.method	Sequência	Basic	Kerberos| Basic	Identifica o método de conexão da origem de dados do banco de dados do Analytic Server.
as.spark.driver.cleanup.delay	Número inteiro	2		A quantidade de tempo (em minutos) após um logout que a JVM do cliente Spark é finalizada.
cleanup.delay	Número inteiro	20		O número de minutos de atraso entre cada execução de limpeza em segundo plano (por exemplo, arquivos de projeto).

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
default.project.versions.tokeep	Número inteiro	25		O número de versões do projeto a serem mantidas ao limpar versões mais antigas do projeto.
distrib.fs.root	Sequência	/user/as_user/analytic-root		A pasta base do Analytic Server para o sistema de arquivos distribuído.
hive.precheckPermission	Booleano	VERDADEIRO		Quando configurado como **TRUE**, o Analytic Server verifica as permissões de arquivo do HDFS para validar os direitos de acesso do usuário para o local de dados da tabela de banco de dados.
hive.sql.check	Booleano	FALSE		Inclui o prefixo `EXPLAIN` às instruções SQL geradas.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
io.sort.mb	Número inteiro	10		A quantidade total de memória de buffer a ser usada (em megabytes) ao classificar arquivos. Por padrão, cada fluxo de mesclagem é alocado 1 MB, o que deve minimizar as buscas. Para obter mais informações, consulte http://hadooptutorial.info/hadoop-performance-tuning/.
java.security.krb5.conf	Sequência			O local do arquivo `krb5.conf` do Kerberos.
jndi.aedb.driver	Sequência			A classe do driver de metastore do Analytic Server.
jndi.aedb.password	Sequência			A senha do metastore do Analytic Server.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
jndi.aedb.url	Sequência			A string de conexão JDBC do repositório de metastore do Analytic Server.
jndi.aedb.username	Sequência			O nome do usuário de metastore do Analytic Server.
join.small.data.size	Número inteiro	1048576		A quantidade máxima de dados (em bytes) que o mecanismo analítico tentará unir em um algoritmo do lado do mapa.
mapred.child.java.opts	Sequência	"-server"		Controla os tamanhos de heap da JVM para mapa e reduz as tarefas executadas no Hadoop. Configure isto como um valor tão grande quanto os nós no cluster possam manipular.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
max.asl.size	Número inteiro	20971520		Tamanho máximo permitido (em bytes) para o programa ASL.
max.datamodel.size	Número inteiro	20971520		Tamanho máximo permitido (em bytes) para a sequência de XML do modelo de dados.
mmr.taskparallel.targets.threshold	Número inteiro	100		As tarefas são processadas pelo M3R quando a razão de destinos/núcleos é menor que este limite. .
mmr.threads	Número inteiro	4		O número de encadeamentos a ser usado para tarefas de redução de mapa na memória (M3R). (*2)

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
mmr.upper.bound.threshold	Numéricos	100		A quantidade máxima de dados que serão processados pelo M3R. Quantidades maiores de dados são processadas pelo Hadoop ou pelo Spark.
nested.groups	Sequência	ativada	enabled| disabled| null (não definido)	Significa que grupos aninhados estão sendo usados no LDAP.
node.max.jobs	Número inteiro	50		O número máximo de tarefas que podem ser executadas simultaneamente em um membro de cluster do Analytic Server.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
orchestrator.thread.pool	Número inteiro	30		O tamanho do orquestrador do conjunto de encadeamentos usado ao enviar o trabalho do mecanismo analítico (por exemplo, **EngineCommands**).
orchestrator.thread.pool.fixed	Booleano	VERDADEIRO		Especifica se o conjunto de encadeamentos do orquestrador é elástico ou fixo.
preferred.mapreduce	Sequência	spark	m3r|hadoop|spark	Define qual mecanismo de redução de mapa preferencial usar.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
resource.pool.default(*1)	Sequência			Configura o valor de **spark.scheduler.pool** quando nenhum mapeamento de consumidor é localizado no **resource.pool.mapping**. Para obter mais informações, consulte https://spark.apache.org/docs/latest/job-scheduling.html.
resource.pool.enabled	Booleano	FALSE		Ativa o uso de definições de mapeamento de fila YARN customizadas (**yarn.queue.mapping**).
resource.pool.mapping(*1)	Mapa		(a:b,c:d....) Yem que a e c são nomes de consumidores do AS e b e d são nomes de conjuntos de recursos do YARN	Mapeia nomes de consumidores do Analytic Server para nomes de conjuntos de recursos do YARN.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
session.max.inactivity.time	Número inteiro	14400		O valor do tempo limite de sessão HTTP (em segundos). O padrão é 14.400 segundos (4 horas).
spark.cache	Booleano	VERDADEIRO		Determina se os RDDs do Spark em cache são utilizados na JVM do cliente do Spark. Por motivos de desempenho, deve ser deixado com o valor padrão de **TRUE**.
spark.dependency.exclude.regex	Sequência		Uma expressão regular válida para excluir arquivos *.jar do caminho de classe da JVM do cliente Spark	Exclui arquivos *.jar problemáticos do caminho de classe da JVM do cliente Spark.
spark.version	Sequência	2.x		A versão do Spark na pilha HDP/CDH que o Analytic Server usa para executar tarefas do Spark.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)				
Nome da Propriedade	**Type**	**Valor-padrão**	**Valores permitidos**	**Descrição**
split.sort.mb	Número inteiro	100		Configura o valor de **io.sort.mb**. Para obter mais informações, consulte https://hadoop.apache.org/docs/r2.4.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml.
yarn.queue.default	Sequência	padrão		O nome da fila YARN padrão que é usado quando nenhum mapeamento válido é localizado em **yarn.queue.mapping**.

Tabela 1. Propriedades customizadas do Analytic Server (continuação)

Nome da Propriedade	Type	Valor-padrão	Valores permitidos	Descrição
yarn.queue.mapping	Mapa		(a:b,c:d....) em que a e c são os nomes do consumidor ou usuário do Analytic Server (conforme determinado por **yarn.queue.mode**) e b e d são nomes das filas do YARN	Mapeia um nome (um usuário ou consumidor) para uma fila do YARN.
yarn.queue.mode	Sequência		user|tenant	Determina se **userName** ou **consumerName** são usados para **mapreduce.job.queuename**. Para obter mais informações, consulte https://hadoop.apache.org/docs/r2.4.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml.

Capítulo 5. Melhores práticas e recomendações do IBM SPSS Analytic Server

As seções a seguir fornecem as melhores práticas e recomendações do Analytic Server relativas às origens de dados, à configuração de cluster e aos fluxos do IBM SPSS Modeler.

Origens de dados

O Analytic Server suporta os tipos de origem de dados a seguir:

- Origens de dados baseadas em arquivo, como arquivos delimitados, de texto fixo e Microsoft Excel.
- Bancos de dados relacionais, tais como Db2, Oracle, Microsoft SQL Server, Teradata, Postgres, Netezza, MySQL e Amazon Redshift.
- Origens de dados do Hive/HCatalog que incluem todos os tipos de dados integrados (por exemplo, ORC e Parquet), bem como qualquer tipo de dados customizados para o qual a implementação apropriada do Serializador-Desserializador do Hive estiver disponível. Além disso, o Analytic Server poderá ser configurado para acessar bancos de dados do NoSQL como HBase, MongoDB, Accumulo, Cassandra, Oracle NoSQL e outros bancos de dados para os quais a implementação apropriada do Manipulador de armazenamento do Hive estiver disponível.

  Nota: O suporte Parquet é limitado à leitura e anexação de tabelas Hive. Quando a substituição das informações da tabela é necessária, uma nova tabela é criada porque o processo de substituição pode resultar na mudança dos valores dos dados, bem como do modelo de dados.
- Origens de dados de tipo geoespacial (baseadas em arquivo de forma e serviços de mapa).

Limitações do Analytic Server em origens de dados Hive/HCatalog

- Se o retrocesso do Hive for necessário para o Nó de Seleção do SPSS Modeler, a expressão de filtragem poderá referenciar apenas colunas particionadas de tipo STRING. Iniciando com o Analytic Server 3.0, foi incluído suporte de tipo de dados para as colunas particionadas a seguir: TINYINT, SMALLINT, INT, BIGINT. A expressão de filtragem estática que é especificada para a origem de dados do Hive pode ter expressões de filtragem para colunas particionadas de qualquer tipo de dados.
- O Analytic Server não suporta origens de dados do HCatalog que são baseadas em visualizações do Hive. Todos os outros tipos de origem de dados (como o Hive SQL) que são baseados em visualizações do Hive são suportados.

Origens de dados do Hive com muitos dados no Cloudera

A ativação do Apache Spark é recomendada para acelerar o processamento de origens de dados do Hive com muitos dados no Cloudera.

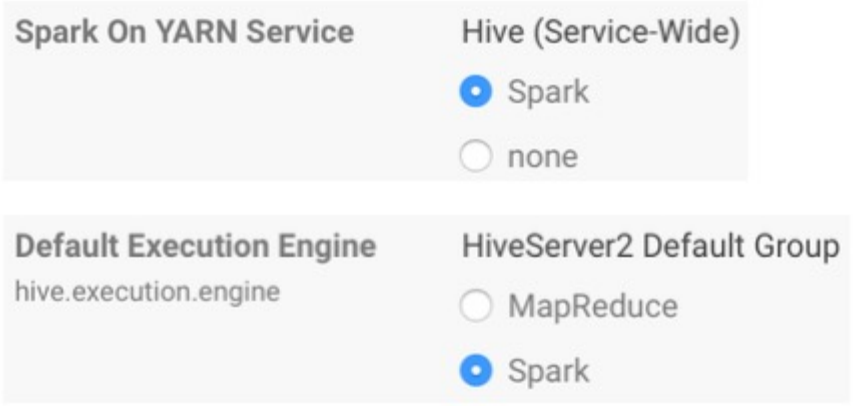

Figura 1. Configurações do Spark

Configuração de cluster - segurança

Personificação do Kerberos

Antes da versão 3.0.1, as instâncias do Analytic Server utilizavam um Nome do principal de usuário na keytab do Analytic Server para autenticar as operações do HDFS quando a segurança do Kerberos estivesse ativada. Iniciando com a versão 3.0.1, o Analytic Server utiliza um Nome do principal de serviço na keytab do Analytic Server juntamente com o nome do usuário solicitante (do usuário que faz a solicitação de pausa) para autenticar as operações do HDFS que utilizem a personificação do Kerberos. O Analytic Server 3.0.1 ou superior deverá incluir atributos de configuração de personificação no HDFS (ou as configurações de serviço do Hive) ao ser executado em um cluster ativado por Kerberos. No caso do HDFS, as propriedades a seguir devem ser incluídas no arquivo `core-site.xml` do HDFS:

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

em que `<analytic_server_service_principal_name>` é o valor padrão `as_user` especificado no campo value `Analytic_Server_User` da configuração do Analytic Server.

As propriedades a seguir também devem ser incluídas no arquivo `core-site.xml` do HDFS em casos nos quais os dados são acessados no HDFS por meio do Hive/HCatalog:

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

Autenticação cross-realm do Kerberos

O Analytic Server suporta autenticação cross-realm do Kerberos. Para ativar esse recurso, deve-se, primeiramente, assegurar que a autenticação cross-realm do KDC esteja ativada e, em seguida, incluir a configuração a seguir na seção **Custom analytics.cfg** da configuração do Ambari do Analytic Server:

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

Configuração de cluster - configurações e resultados de ajuste de desempenho

Configuração do Spark

O Analytic Server usa o modo `yarn-client` para interagir com YARN e executar tarefas Spark no cluster Hadoop.

Configuração customizada do Analytic Server:

- As configurações do Ambari são definidas na seção **Custom analytics.cfg** da configuração do Ambari do Analytic Server.
- As configurações do Cloudera estão localizadas na seção **Fragmento de Configuração Avançada do Servidor Analítico (Válvula de Segurança) para analyticserver-conf/config.properties** do Gerenciador do Cloudera.

1. Considere aumentar o valor da definição de configuração **spark.driver.memory** incluindo um item de configuração na configuração customizada do Analytic Server (quando não configurado explicitamente, o valor padrão é `1g`). Por exemplo:

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

2. Selecione a partir de um dos seguintes Analytic Server com opções de uso de recurso do Spark.

   - **Opção A: configuração de alocação de recurso estático**

     Existem 3 parâmetros que devem ser configurados na configuração customizada do Analytic Server:

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

     As etapas a seguir descrevem como determinar os valores de parâmetro.

a. Estabeleça a porcentagem, em termos de CPU e memória, que o Analytic Server pode alocar permanentemente para o Spark. Isso resulta em um número específico de núcleos (C) e uma quantia fixa de memória que podem ser usados em cada máquina (M).

b. Estabeleça o número de executores (E) que cada máquina pode executar. Esses executores são executados como contêineres (processos) do Hadoop separados em cada nó do cluster. Geralmente, um valor maior que 2 é apropriado, mas o valor deve ser menor que o número total de núcleos. A memória que é alocada para o Spark é dividida entre esses executores, de maneira que selecionar um valor alto para esse parâmetro diminuirá a quantia de memória que será alocada para cada contêiner.

c. Estabeleça o número de núcleos que são usados por cada executor (CE). Geralmente, esse valor é C/E (o número de núcleos de cada máquina que são alocados para o aplicativo Spark, dividido pelo número total de executores).

d. Estabeleça a quantia de memória que é usada para cada executor (ME). Isso geralmente é M/E.

Nota: O número de executores e núcleos usados deve ser balanceado de forma que a quantidade de cada memória do executor seja maior que 3G * CE. Cada núcleo de cada executor deve possuir alocados pelo menos 3G de memória, que serão usados como memória de armazenamento ou de computação.

```
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
```

Figura 2. Configurações do Spark de custom analytics.cfg

- **Opção B: configuração de alocação de recurso dinâmico**

  Ao usar essa opção, todos os executores alocados pelo YARN são aumentados/diminuídos dinamicamente, de acordo com os recursos reais disponíveis do cluster inteiro.

  A configuração mínima é:

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

  Uma configuração típica é:

```
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
```

  Notas:

  – `spark.executor.instances = <E>` não deve ser usado, caso contrário, a alocação de recurso estático é empregada.

  – As considerações com relação aos núcleos e aos valores de memória do executor são idênticas àquelas discutidas na Opção A.

3. É possível desativar o cache do Spark na configuração customizada do Analytic Server usando as configurações a seguir:

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

Figura 3. Configurações de cache do Spark de custom analytics.cfg

O cache do Spark não deverá ser desativado quando grandes fluxos do IBM SPSS Modeler forem usados. A desativação do cache do Spark nessa instância resulta em fluxos de execução mais lentos, mas evita condições de falta de memória que podem ocorrer quando a quantidade especificada de memória por executor é pequena.

Configuração JVM

Configurações Ambari:

1. Na configuração do Ambari do Analytic Server, configure a quantidade de memória que o servidor pode usar para o processamento local. O valor padrão (2 GB) pode ser usado seguramente para fluxos pequenos a médios, mas um tamanho de heap de valor mais alto (por exemplo, 10 GB) deve ser usado para maiores fluxos.

   Servidor analítico > **Configuração** > **analytic-jvm-options avançadas**
2. Substitua `-Xmx2048M` por `-Xmx10G`, salve a configuração e reinicie o Analytic Server.

Figura 4. Configurações de analytic-jvm-options avançadas

Configurações Cloudera:

1. No Cloudera Manager, navegue para a guia **Configuração** do serviço Analytic Server e atualize o controle `jvm-options` para configurar a quantidade de memória que o servidor pode usar para processamento local. O valor padrão (2 GB) pode ser usado seguramente para fluxos pequenos a médios, mas um tamanho de heap de valor mais alto (por exemplo, 10 GB) deve ser usado para maiores fluxos.

   Serviço do Analytic Server > **Configuração** > **jvm-options**
2. Substitua `-Xmx2048M` por `-Xmx10G`, salve a configuração e reinicie o Analytic Server.

Configuração do YARN MapReduce2:

- Se você deve executar tarefas MapReduce em paralelo com tarefas Spark para execução do Analytic Server, o cluster do YARN deve ser configurado para ter pelo menos 4 GB de memória por cada contêiner do YARN.

Configuração do Zookeeper:

- Cloudera requer que você atualize manualmente a configuração do Zookeeper. Para obter mais informações, consulte https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_zookeeper_server_maintain.html.
- Se você usar fluxos complexos do SPSS Modeler ou dados amplos (um grande número de campos), poderá ter problemas com tarefas falhas em virtude de uma conexão de Analytic Server–Zookeeper interrompida. O problema é o resultado do grande tamanho do programa que o Servidor SPSS Modeler envia ao Analytic Server. É menos provável que o problema ocorra no Analytic Server 3.0 (ou superior). Use as etapas a seguir para resolver o problema:

  1. No console do Ambari, navegue para a guia **Configurações** do serviço do Zookeeper, inclua a linha a seguir no modelo `zookeeper-env` sob **zookeeper-env avançado** e, em seguida, reinicie o serviço do Zookeeper.

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

Figura 5. Configurações de modelo do zookeeper-env

2. No console do Ambari, navegue para a guia **Configurações** do serviço do Analytic Server, inclua o seguinte em **analytics-jvm-options avançadas** e, em seguida, reinicie o serviço do Analytic Server.

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

Figura 6. Configurações de analytics-jvm-options avançadas

Nota: Se o problema persistir, aumente o valor `-Djute.maxbuffer` de 2097152 para 4194304 em ambos os lugares.

Recomendações de fluxo do IBM SPSS Modeler

Nota: A maioria das recomendações a seguir também se aplica a dados pequenos.

Protótipo em dados pequenos

Ao ter um experimento com um fluxo, você, muitas vezes, inclui alguns nós, testa o fluxo até esse ponto, talvez inclua um nó para verificar a saída tabular ou gráfica e, em seguida, continua o fluxo. Normalmente, você não pode se dar ao luxo de fazer uma passagem de dados de big data toda vez que testar o fluxo.

Criar uma amostra de dados adequada de big data permite que você teste o fluxo com relação aos dados reais sem incorrer em penalidade de tempo que é necessária ao executar uma passagem de dados completa. A amostra de dados deve conter dados suficientes para a execução bem-sucedida do fluxo. Por exemplo, se você planeja analisar transações em lojas que estiverem localizadas em Minnesota, a sua amostra de dados deverá conter transações de lojas em Minnesota.

Após a amostragem, será possível:

- Criar um cache da amostra de dados no cluster no qual o big data reside ou

  Prós -Simples e não requer comutação de nós de origem

  Contras - O cache desaparecerá quando a sessão terminar
- Criar uma nova origem de dados do Analytic Server que contenha a amostra de dados ou

  Prós - Origem de dados permanente

  Contras -Requer a edição/comutação de nós de origem
- Fazer download da amostra de dados em seu sistema local e criar uma origem de dados local

  Prós -Não consome recursos de cluster ao realizar protótipo; o cliente do SPSS Modeler é mais eficiente que o Analytic Server quando você trabalha com dados pequenos.

  Contras - Requer a comutação de nós de origem

Crie nós de Tipo e Filtro separados dos nós de origem

Cada nó de origem do SPSS Modeler também tem a funcionalidade combinada dos nós de Filtro e Tipo. Isso é útil para manter a tela aperfeiçoada, mas torna difícil quando você alterna para Tipos de nó de

origem diferentes. Além disso, isso obscurece o fato de que as operações de Tipo e Filtro estão ocorrendo.

Coloque nós de Filtro e de Seleção o mais próximo possível do Nó de origem

Isso reduz o número de registros em operações de recebimento de dados.

Evite o nó de Classificação sempre que possível

O Analytic Server não suporta as otimizações em nós que dependem dos dados que estiverem sendo classificados (como o nó de Mesclagem). Desse modo, um nó de Classificação intermediário raramente está fazendo algo útil. O nó de Classificação tem valor quando seguido imediatamente por um nó de Amostra para obter os registros de N Superior (ou N Inferior).

Calcule apenas os campos que serão usados

Não calcule um campo e, em seguida, filtre-o imediatamente.

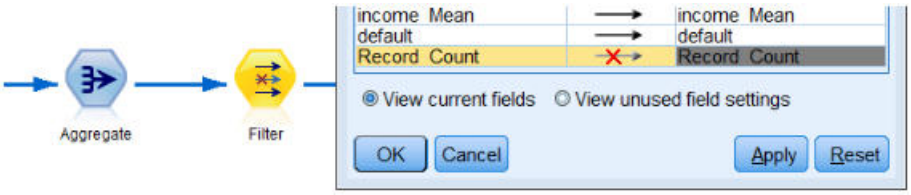

Figura 7. Opções de campo do modelador

Sempre que possível, sem tornar as expressões difíceis de entender, evite criar campos numerosos, temporários. Por exemplo, em vez de definir o exemplo a seguir:

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

defina o exemplo a seguir, como alternativa:

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

Compactar expressões provisórias em sequenciais desta forma poderá fazer aumentar o desempenho quando um grande número de campos for transformado.

Configure o armazenamento nos dados de origem

Operações que mudam um intermediário de tipo de armazenamento do campo (por exemplo, sequência para número inteiro) podem ser prejudiciais para o desempenho geral. É possível configurar o armazenamento para campos, ao definir origens de dados no Console do Analytic Server, para evitar repetir essas conversões.

Use o SPSS Modeler ao trabalhar com dados pequenos

Manipule big data com o Analytic Server e, em seguida, use o SPSS Modeler para concluir os cálculos em dados pequenos.

Selecione as propriedades de fluxo relacionadas ao Analytic Server

Configure as propriedades de fluxo relevantes (**Ferramentas** > **Opções** > **Propriedades de fluxo** > **Servidor analítico**) e decida se deve-se permitir o processamento de dados para sair do Analytic Server e continuar no SPSS Modeler (quando um nó não puder ser executado no Analytic Server).

Por padrão, o SPSS Modeler é configurado para relatar um erro e parar a execução nesta situação. É possível efetuar bypass do erro mudando a configuração de `Error` para `Warn` e ajustando o limite de quantos dados podem ser processados no SPSS Modeler. Por exemplo, é possível atualizar a taxa de transferência de dados do valor de registro padrão 10.000 (se necessário). Observe que esse limite também se aplica ao visualizar resultados que usam o nó de tabela do SPSS Modeler. Se o limite for excedido, o SPSS Modeler relata que a Busca de dados excedeu o limite configurado nas propriedades de fluxo.

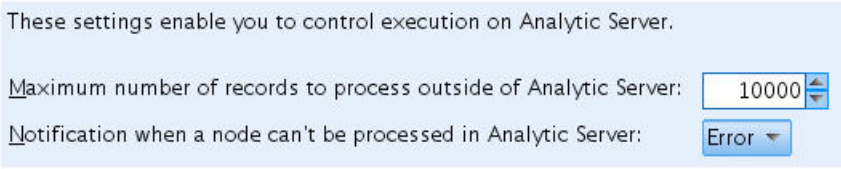

Figura 8. Configurações do Servidor analítico

Use os nós de origem do Analytic Server

O Analytic Server pode se conectar a diferentes origens de dados do banco de dados, mas o SPSS Modeler requer que todos os Nós de origem sejam Nós de origem do Analytic Server (em ordem para o fluxo inteiro ser executado como uma tarefa do Analytic Server). Para o fluxo inteiro ser executado no Analytic Server, o nó de origem do banco de dados deve ser mudado para um Nó de origem do Analytic Server e uma origem de dados do banco de dados do Analytic Server deve ser criada no Console do Analytic Server.

Considere como os nós não suportados são usados

O Analytic Server não suporta todos os nós (o nó de Transposição é um bom exemplo). Para mesclar os resultados de uma operação de transposição com o restante do fluxo e executá-la no Analytic Server, um subfluxo que inclui um nó de Transposição deve ser gravado em uma origem de dados do Analytic Server que use um nó de Exportação do Analytic Server. É possível, então, conectar um Nó de origem do Analytic Server no qual o fluxo foi interrompido para gravar no Analytic Server.

Nota: A operação de transposição é adequada para operações únicas ou raramente executadas, mas não deve ser usada para operações de fluxo de rotina.

Determine se um fluxo funcionará no Analytic Server antes que ele seja executado

Após você preparar um fluxo para execução no Analytic Server, selecione um nó terminal e use o recurso de visualização do SPSS Modeler (o controle **Visualizar execução** na barra de ferramentas) para verificar se qualquer nó que estiver envolvido na execução do nó terminal funcionará no Analytic Server (sem executar o fluxo). Os problemas são relatados na janela de mensagens.

Combine operações back-to-back de Mesclagem

Combine uma série de nós de Mesclagem em um único nó quando eles tiverem as mesmas chaves e o tipo de junção.

Combine subfluxos idênticos

Tente combinar subfluxos idênticos sempre que possível, especialmente se eles contiverem operações dispendiosas (por exemplo, mesclagem e classificação). O SPSS Modeler executa essas operações uma vez e usa cache para melhorar o desempenho. No exemplo a seguir, os fluxos são idênticos até o nó **newField**.

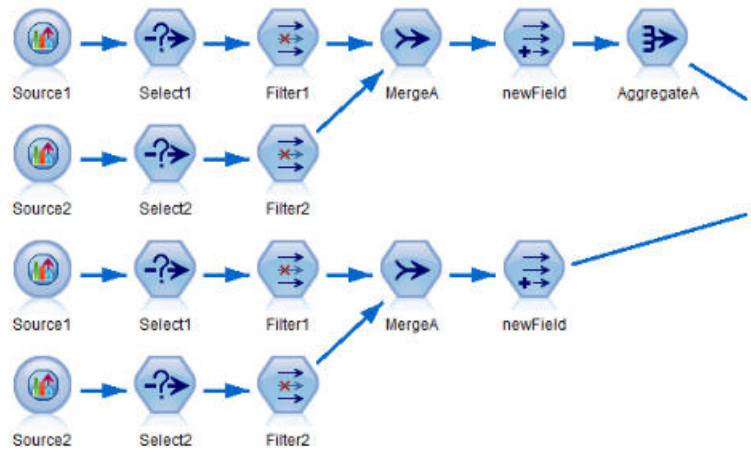

Figura 9. Fluxo de exemplo

Será mais eficiente (e fácil de manter) se o subfluxo, como alternativa, for estruturado como segue:

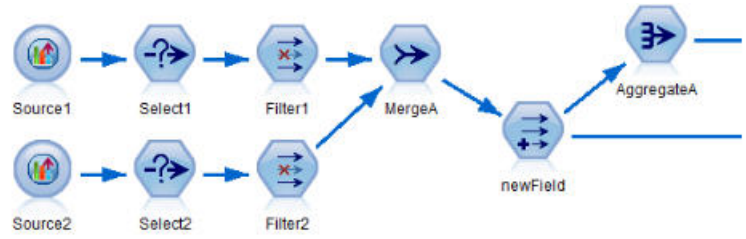

Figura 10. Fluxo de exemplo

Remova nós Tipo extras

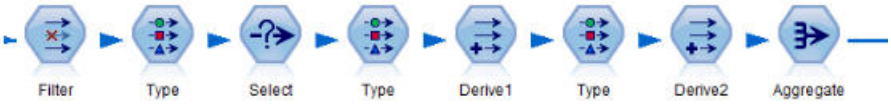

Figura 11. Fluxo de exemplo

Evite nós Tipo desnecessários ao executar com relação ao Analytic Server. A operação `Read Values` do nó Tipo inicia uma tarefa do MapReduce. Normalmente, isso é uma economia única, a menos que você limpe os valores de nó Tipo.

Documente completamente cada fluxo

O exemplo a seguir mostra um fluxo complexo que contém uma série de subfluxos.

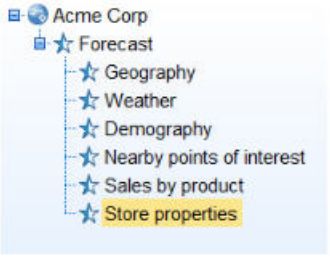

Figura 12. Exemplo de subfluxo

Nesses casos, é importante nomear adequadamente os supernós e documentar o fluxo (como você documentaria o código). Um comentário claro pode fornecer informações valiosas para outros analistas que leem ou mantêm o fluxo. Por exemplo:

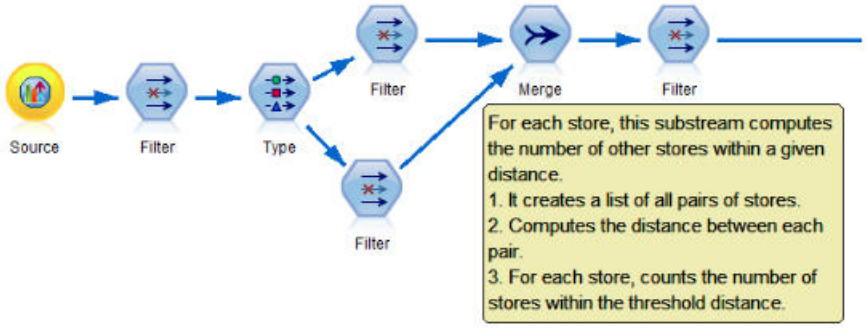

Figura 13. Exemplo de fluxo com comentários

Ao desenvolver fluxos, use caches do SPSS Modeler para armazenar rapidamente os resultados intermediários

Em fluxos que forem executados com relação ao Analytic Server, o armazenamento em cache funciona armazenando os dados em uma parte específica do fluxo para arquivos temporários no HDFS (em

oposição a armazenar no servidor SPSS Modeler). Os caches trabalham bem com big data e são seguros para usar em fluxos que são executados no Analytic Server.

Capítulo 6. Resolução de problemas

O Analytic Server fornece várias ferramentas úteis para a determinação de problema.

Criação de log

O Analytic Server cria arquivos de log do cliente e arquivos de rastreio que são úteis para diagnosticar problemas. Com a instalação padrão do Liberty, é possível localizar os arquivos de log no diretório `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/logs`.

A configuração padrão da criação de log produz dois arquivos de log que são substituídos em uma base diária.

as.log
Este arquivo contém o resumo de alto nível de mensagens de aviso e erro informativas. Verifique esse arquivo primeiro quando ocorrerem erros do servidor que não podem ser resolvidos usando a mensagem de erro que é exibida na Interface com o usuário.

as_trace.log
Este arquivo contém todas as entradas de `ae.log`, mas inclui mais informações que são primariamente voltadas ao suporte IBM e ao desenvolvimento para fins de depuração.

Analytic Server usa Apache LOG4J como seu recurso de criação de log subjacente. Usando LOG4J, a criação de log pode ser dinamicamente ajustada editando o arquivo de configuração `{AS_SERVER_ROOT}/configuration/log4j.xml`. O Suporte pode solicitar que faça isso para ajudar a diagnosticar problemas ou você pode querer modificar isto para limitar o número de arquivos de log mantidos. Mudanças no arquivo são automaticamente detectadas em poucos segundos, assim o Analytic Server não precisa ser reiniciado.

Para obter mais informações sobre log4j e o arquivo de configuração, consulte a documentação no website Apache oficial em http://logging.apache.org/log4j/.

Informações de versão

É possível determinar que versão do Analytic Server é instalada verificando a pasta `{AS_ROOT}/properties/version`. Os arquivos a seguir contêm informações da versão.

IBM_SPSS_Analytic_Server-*.swtag
Contém informações detalhadas do produto.

version.txt
Versão e número da construção para o produto instalado.

Coletor de logs

Quando os problemas não podem ser resolvidos diretamente revisando os arquivos de log, é possível compactar todos os logs e enviá-los ao suporte IBM. Há um utilitário que é fornecido para tornar mais simples a coleta de todos os dados necessários.

Usando um shell de comando, execute os comandos a seguir:

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

Esses comandos criam um arquivo compactado em `{AS_ROOT}/bin`. O arquivo compactado contém todos os arquivos de log e informações de versão do produto.

Problemas Comuns

Esta seção descreve alguns problemas de administração comuns e como é possível corrigi-los.

Fluxos em execução

Tarefas R convertem palavras que não estão em inglês para Unicode

Em clusters do Cloudera, se a codificação do sistema de servidores Hadoop não for UTF-8, R converte as palavras que não estão em inglês para Unicode.

1. Navegue para a guia de configuração do YARN no console do Gerenciador do Cloudera.
2. Inclua as configurações a seguir no campo "Fragmento de Configuração Avançada de Ambiente do NodeManager (Válvula de Segurança)".

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

A execução das tarefas PySpark falhou

Certifique-se de que o serviço Spark esteja implementado em todos os nós do Analytic Server e em todos os gerenciadores de nós.

As tarefas PySpark falharam ao serem executadas em ambientes ativados do Kerberos

Deve-se executar o comando `kinit` e, em seguida, reiniciar o Analytic Server antes que os testes do PySpark sejam executados com sucesso. Por exemplo:

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
```

O nó XGBoost-AS falha ao executar

Você pode enfrentar o erro a seguir ao tentar executar o nó XGBoost-AS:

```
Error executing ASL: Execution failed. Reason: ml.dmlc.xgboost4j.java.XGBoostError:
XGBoostModel training failed
```

Configuração customizada do Analytic Server

1. Aplique as definições de configuração customizada do Analytic Server a seguir para resolver o problema.

```
spark.executor.memory=12g (or above)
```

   - As configurações do Ambari são definidas na seção de configuração **Customizar analytics.cfg** do Ambari.
   - As configurações do Cloudera estão localizadas na seção **Fragmento de Configuração Avançada do Servidor Analítico (Válvula de Segurança) para analyticserver-conf/config.properties** do Gerenciador do Cloudera.

2. Reinicie o serviço do Analytic Server após atualizar a configuração customizada do Analytic Server.

Configuração do IBM SPSS Modeler

Ative a configuração do nó XGBoost-AS IBM SPSS Modeler **Opções de Compilação** > **Geral** > **Usar Memória Externa**.

Erros de memória

Configurando YARN após erros de memória do executor

O erro a seguir poderá ocorrer quando a memória do executor necessária estiver acima do limite máximo:

```
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'.
```

As etapas a seguir fornecem as definições de configuração YARN que são necessárias para resolver o problema.

Para Ambari

1. Na interface com o usuário Ambari, acesse **YARN** > **Configurações** > **Definições**.
2. Aumente o nó de **memória (a memória que está alocada para todos os contêineres YARN)** para 8192 MB.
3. Aumente os valores de contêiner
   - **Tamanho Mínimo do Contêiner (Memória)** para 682 MB
   - **Tamanho Máximo do Contêiner (Memória)** para 8192 MB
4. Aumente o **Tamanho Máximo do Contêiner (VCores)** para 3.
5. Reinicie o YARN, Spark e o serviço do Analytic Server.

Para Cloudera

1. Aumente o `yarn.nodemanager.resource.memory-mb` para 8 GB
   - Na interface com o usuário do Cloudera Manager, acesse **Serviço do YARN** > **Configurações** > **Memória de contêiner de procura** e aumente o valor para 8GB.
2. Na interface com o usuário do Cloudera Manager, acesse **Serviço YARN** > **Links Rápidos** e selecione **Conjuntos de Recursos Dinâmicos**.
3. Em **Configuração**, clique em **editar** para cada um dos conjuntos disponíveis e, em **YARN**, configure o valor de **Máx. de Aplicativos em Execução** para 4.
4. Reinicie o YARN, Spark e o serviço do Analytic Server.

Memória do executor necessária

O erro a seguir pode ocorrer quando a memória do executor necessária está acima da configuração atual:

```
AS Executor Removed:SparkListenerExecutorRemoved(1557300399468,716,Container killed by YARN
for exceeding memory limits.
2.0 GB of 2 GB physical memory used. Consider boosting spark.yarn.executor.memoryOverhead.)
```

1. Aplique as definições de configuração customizada do Analytic Server a seguir para resolver o problema.

   ```
   spark.executor.memory=4g (or above)
   ```

   - As configurações do Ambari são definidas na seção de configuração **Customizar analytics.cfg** do Ambari.
   - As configurações do Cloudera estão localizadas na seção **Fragmento de Configuração Avançada do Servidor Analítico (Válvula de Segurança) para analyticserver-conf/config.properties** do Gerenciador do Cloudera.
2. Reinicie o serviço do Analytic Server após atualizar a configuração customizada do Analytic Server.

Hadoop com Apache Spark 2.2

- A maioria das tarefas `forcespark` e `forcehadoop` falham quando o Hadoop e o Apache Spark 2.2 existem no mesmo ambiente. O erro aparece no log do aplicativo do YARN como: `java.lang.NoClassDefFoundError: org/apache/hadoop/fs/FSDataInputStream`.

O problema pode ser resolvido editando manualmente o arquivo `/etc/spark2/conf/spark-defaults.conf` da seguinte forma:

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

- Quando duas versões do JDK são instaladas no mesmo sistema, o Cloudera usa o JDK 1.7 enquanto o Spark 2.2 usa o JDK 1.8. A execução das tarefas `forcespark` ou `forcehadoop` com o Apache Spark 2.x pode resultar em falha em todas as tarefas com a mensagem de erro a seguir:

  ```
  A execução falhou. Motivo: org/apache/spark/api/java/function/PairFunction : versão 52.0 de major.minor não suportada
  ```

  Para o Cloudera, inclua a seguinte linha na seção **Fragmento de Configuração Avançada do Servidor Analítico (Válvula de Segurança) para server.env** do Gerenciador do Cloudera:

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

Concedendo a autoridade do admin a usuários do Apache Hive UDF

É possível encontrar um erro `Invalid function` após o registro do Apache Hive UDF do Analytic Server. Por padrão, há duas funções Hive (`admin` e `public`). Usuários Hive pertencem à função `public`. O Hive UDF requer que os usuários registrados possuam o privilégio `admin` (a segurança do Hive está ativada).

Para conceder a autoridade do `admin` para usuários do Hive UDF:

1. Efetue login no Beeline como Hive:

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

2. Execute o comando a seguir no Beeline:

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

   Nota: Outros comandos SQL úteis incluem:

   Mostrar as funções já designadas ao usuário `hive`

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

   Mostrar quais usuários estão designados para a função `public`

   ```
   show principals public;
   ```

3. Reiniciar o Hive e registrar novamente o 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
   ```

Erro do HiveDB

É possível encontrar o erro a seguir ao gravar em um HiveDB:

```
(AEQAE4805E) Falha na execução. Razão: com.google.common.io.Closeables.closeQuietly(Ljava/io/Closeable;)
```

O erro é causado por diversas versões do arquivo `guava-*.jar` no cluster Hadoop. O erro pode ser resolvido ao executar as etapas a seguir (o exemplo usa o HDP 3.1):

1. Abra o console do Ambari e pare o serviço Analytic Server.
2. Copie o `/usr/hdp/3.1.0.0-78/spark2/jars/guava-14.0.1.jar` para `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib`.
3. No console do Ambari, atualize o serviço Analytic Server e, em seguida, inicie o serviço Analytic Server.

Combinação de origens de dados do Hive e do HCatalog

O Analytic Server não suporta a combinação de origens de dados do Hive e do HCatalog no mesmo fluxo IBM SPSS Modeler.

Ajuste de desempenho

Esta seção descreve maneiras para otimizar o desempenho de seu sistema.

O Analytic Server é um componente da estrutura Ambari que utiliza outros componentes, como o HDFS, o YARN e o Spark. Técnicas de ajuste de desempenho comuns para Hadoop, HDFS e Spark aplicam-se a cargas de trabalho do Analytic Server. Cada carga de trabalho do Analytic Server é diferente; portanto, é necessária a experimentação de ajuste com base em sua carga de trabalho de implementações específica. As propriedades e dicas de ajuste a seguir são mudanças chave que impactaram os resultados dos testes comparativos e de ajuste de escala do Analytic Server.

Quando a primeira tarefa for executada no Analytic Server, o servidor iniciará um aplicativo Spark persistente que ficará ativo até que o Analytic Server seja encerrado. O aplicativo Spark persistente irá alocar e manter todos os recursos de cluster alocados para ele pela duração da execução do Analytic Server, mesmo se uma tarefa do Analytic Server não estiver ativamente em execução. Deve-se pensar com cuidado na quantia de recursos alocados para o aplicativo Spark do Analytic Server. Se todos os recursos de cluster forem alocados para o aplicativo Spark do Analytic Server, então, outras tarefas poderiam ser atrasadas ou não ser executadas. Essas tarefas poderiam ser enfileiradas esperando por suficientes recursos livres e esses recursos serão consumidos pelo aplicativo Spark do Analytic Server.

Se múltiplos serviços do Servidor Analítico forem configurados e implementados, cada instância de serviço poderia alocar potencialmente o seu próprio aplicativo Spark persistente. Por exemplo, se dois serviços do Servidor Analítico forem implementados para suportar failover de alta disponibilidade, então, você poderia ver dois aplicativos Spark persistentes ativos, cada um alocando recursos de cluster.

Uma complexidade adicional é que, em determinadas situações, o Servidor Analítico pode iniciar uma tarefa de redução de mapa que irá requerer recursos de cluster. Essas tarefas de redução de mapa irão requerer recursos que não estão alocados para o aplicativo Spark. Os componentes específicos que requerem tarefas de redução de mapa são construções de modelo de PSM.

As propriedades a seguir podem ser configuradas para alocar recursos ao aplicativo Spark. Se elas forem configuradas no spark-defaults.conf da instalação do Spark, então, elas serão alocadas para todas as tarefas do Spark executadas no ambiente. Se elas forem configuradas na configuração do Servidor Analítico como propriedades customizadas sob a seção “Analytic.cfg customizada”, então, elas serão alocadas apenas para o aplicativo Spark do Servidor Analítico.

spark.executor.memory

Quantia de memória a ser usada por processo de executor.

spark.executor.instances

O número de processos de executor a serem iniciados.

spark.executor.cores

O número de encadeamentos do trabalhador do executor por processo de executor. Esse valor deve ser entre 1 e 5.

Um exemplo de configuração das três propriedades chave do Spark. Há 10 nós de dados em um cluster de HDFS e cada nó de dados tem 24 núcleos lógicos e 48 GB de memória e está executando somente processos do HDFS. Aqui está uma maneira para configurar as propriedades para esse ambiente, supondo que você esteja executando apenas tarefas do Servidor Analítico nesse ambiente e queira alocação máxima para um único aplicativo Spark do Servidor Analítico.

- Set spark.executor.instances=20. Isso tentaria executar 2 processos de executor do Spark por nó de dados.

- Set spark.executor.memory=22G. Isso iria configurar o tamanho de heap máximo para cada processo de executor do Spark para 22 GB, alocando 44 GB em cada nó de dados. Outras JVMs e o S.O. precisam de memória extra.
- Set spark.executor.cores=5. Isso fornecerá 5 encadeamentos do trabalhador para cada executor do Spark, para um total de 10 encadeamentos do trabalhador por nó de dados.

Monitorar a UI do Spark para executar tarefas

Se você vir Spill para disco que poderia impactar o desempenho. Algumas possíveis soluções são:

- Aumentar a memória e alocá-la para executores do Spark através de **spark.executor.memory**.
- Reduzir o número de **spark.executor.cores**. Isso reduzirá o número de encadeamentos de trabalho simultâneos que alocam memória, mas também reduzirá a quantia de paralelismo para as tarefas.
- Mude as propriedades de memória do Spark. Porcentagem de alocação de **spark.shuffle.memoryFraction** e **spark.storage.memoryFraction** do heap de executor do Spark para Spark.

Assegure-se de que o nó do nome tenha memória suficiente

Se o número de blocos no HDFS for grande e crescente, certifique-se de que o heap de nó de nome aumente para acomodar esse crescimento. Essa é uma recomendação de ajuste do HDFS comum.

Altere a quantia de memória usada para armazenamento em cache

Por padrão, **spark.storage.memoryFraction** tem um valor 0,6. Esse pode ser aumentado até 0,8 em caso do tamanho de bloco de HDFS de dados ser 64 MB. Se o tamanho de bloco de HDFS dos dados de entrada for maior que 64 MB, então, esse valor poderia ser aumentado apenas se a memória alocada por tarefa for maior que 2 GB.

Ajustando o Desempenho de Escoragem de Modelo

É possível melhorar o desempenho de tarefas escoragem de modelo em grandes conjuntos de dados com o mecanismo Apache Spark usando as etapas a seguir. Observe que essas etapas não devem impactar a operação de serviços do Servidor não Analítico no cluster.

1. Verifique se libtcmalloc_minimal.so{/version} já está instalada em cada nó no cluster.

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

2. Se libtcmalloc_minimal.so não estiver instalada, instale o pacote específico do sistema operacional que contém a biblioteca libtcmalloc_minimal em cada nó no em seu cluster ou construa e instale manualmente libtcmalloc_minimal. Por exemplo:

   Ubuntu:

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

   Red Hat Enterprise Linux 6.x (x64):

   a. Instale o repositório do EPEL para RedHat (se ainda não estiver instalado)

      ```
      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`

   Construção Manual:

   a. Faça download do gperftools-2.4.tar.gz no link https://github.com/gperftools/gperftools/releases

   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. Observe um dos locais do arquivo de biblioteca instalado libtcmalloc_minimal.so{.version}, conforme retornado a partir da execução do comando a seguir em um ou mais dos nós.

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

Se o cluster tiver nós executando uma mistura de sistemas operacionais, poderia haver múltiplos locais para esse arquivo.

4. No console do Ambari, acesse a configuração do Servidor Analítico e sob a seção Analytics.cfg customizada, configure o spark.executorEnv.LD_PRELOAD chave usando o local da biblioteca como o valor. Após fazer essa mudança, reinicie o serviço do Servidor Analítico. Por exemplo, se a biblioteca estiver instalada em /usr/lib64/libtcmalloc_minimal.so.4, a configuração seria:

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

Se múltiplos locais forem necessários, use um espaço para separá-los, como no exemplo a seguir.

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

Se algum nó não tiver a biblioteca `libtcmalloc_minimal.so` instalada em um dos locais configurados, isso não causará um erro, mas o desempenho de escoragem de modelo poderá ser mais lento nesse nó.

Junção do lado do mapa do Spark

A implementação de junção do Spark do Analytic Server não suporta a funcionalidade de junção do lado do mapa (A junção do Spark é principalmente um lado de redução). A implementação não se aproveita de junções do lado do mapa para otimizar junções quando uma entrada é pequena. Não se aproveitar da junção do lado do mapa resultará em uma tarefa do Spark extremamente intensiva de recurso que eventualmente falhará.

Para otimizar junções ao executar junções do lado do mapa do Spark do Analytic Server (ou tarefas nativas do Spark que são baseadas no menor tamanho do RDD), será possível incluir a propriedade `spark.msj.maxBroadcast` no arquivo `analytics.cfg` (`SPSS Analytic Server/Configs/Custom analytics.cfg`) ou no `analytics-meta`.

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos nos EUA. Este material pode estar disponível pela IBM em outros idiomas. Entretanto, poderá ser necessário ter uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos neste documento em outros países. Consulte seu representante IBM local para obter informações sobre os produtos e serviços disponíveis atualmente em sua área. Qualquer referência a um produto, programa ou serviço IBM não significa que apenas produtos, programas ou serviços IBM possam ser usados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser usado em substituição. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do usuário.

A IBM pode ter patentes ou solicitações de patentes pendentes relativos a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

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

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS DE NÃO INFRAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Quaisquer referências nessas informações em sites não IBM são fornecidas por conveniência apenas e não de modo a servir como endosso para esses websites. Os materiais contidos nesses websites não fazem parte dos materiais desse produto IBM e a utilização desses websites é de inteira responsabilidade do Cliente.

A IBM pode usar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146

Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriados, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito neste documento e todo o material licenciado disponível são fornecidos pela IBM sob os termos do IBM Customer Agreement, Contrato de Licença do Programa Internacional da IBM ou qualquer contrato equivalente.

Os exemplos de clientes e dados de desempenho citados são apresentados com propósitos meramente ilustrativos. Os resultados reais de desempenho podem variar de acordo com as configurações específicas e condições de operação.

Informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

As declarações relacionadas a direção ou intenção futuros da IBM estão sujeitas a alteração ou retirada sem aviso prévio e representam metas e objetivos apenas.

Todos os preços IBM mostrados são preços de varejo sugeridos pela IBM, são atualizados e estão sujeitos a alterações sem aviso prévio. Os preços dos revendedores podem variar.

Estas informações são apenas para fins de planejamento. As informações nesta publicação estão sujeitas a alterações antes que os produtos descritos se tornem disponíveis.

Estas informações contêm exemplos de dados e relatórios utilizados em operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm exemplos de dados e relatórios utilizados em operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

Cada cópia ou parte desses programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

Marcas Registradas

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em vários países no mundo todo. Outros nomes de empresas, produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas comerciais ou marcas registradas da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

IT Infrastructure Library é uma marca registrada da Central Computer and Telecommunications Agency que agora faz parte do Departamento de Comércio do Governo.

Intel, logotipo Intel, Intel Inside, logotipo Intel Inside, Intel Centrino, logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas registradas da Microsoft Corporation nos Estados Unidos e/ou em outros países.

ITIL é uma marca registrada e uma marca de comunidade registrada da The Minister for the Cabinet Office e está registrada no U.S. Patent and Trademark Office.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

Cell Broadband Engine é uma marca comercial da Sony Computer Entertainment, Inc. nos Estados Unidos e/ou em outros países e é utilizada sob licença.

Linear Tape-Open, LTO, o logotipo LTO, Ultrium e o logotipo Ultrium são marcas comerciais da HP, IBM Corp. e Quantum nos Estados Unidos e em outros países.