第 3 章　面向混合云的资源管理

3.1　混合云环境下面向安全的科学工作流数据布局策略

在当今数字化时代，科学工作流已成为科学研究进程中处理大规模数据的重要手段。随着数据复杂度的不断增加，传统网络和集群环境的资源已不足以满足科学工作流的部署需求。相比之下，混合云计算拥有强大的计算和存储资源，为科学工作流的部署提供了一种理想化平台。然而，现有的混合云环境下的科学工作流数据布局策略通常将隐私数据固定存储，而没有充分考虑数据在分布式环境中的安全性。针对该问题，本节构建了一种混合云环境下面向数据安全的科学工作流数据布局模型，并提出安全等级分级机制，其考虑数据的安全需求和数据中心所能提供的安全服务，旨在确保数据隐私安全的同时最大程度降低数据传输时延。另外还提出了一种基于自适应粒子群优化算法（Adaptive Particle Swarm Optimization Algorithm based on SA and GA，SAGA-PSO）的数据布局策略。SAGA-PSO 基于粒子群优化算法框架，在粒子更新操作中引入遗传算法的交叉和变异算子，提高了种群进化的多样性；在全局最优粒子替换中引入模拟退火算法的 Metropolitan 准则，有效避免算法陷入局部极值。相关实验结果表明，本节提出的基于 SAGA-PSO 的数据布局策略在满足数据安全需求的同时能够有效降低传输时延。

3.1.1　引言

随着云计算技术的广泛应用，越来越多的云服务提供商推出了各种类型的云存储服务，以满足用户对数据存储和管理的需求。云存储服务的出现为科学工作流的执行提供了新的选择，用户可以将数据存储在云端，通过云计算平台进行数据处理和分析，从而提高数据的处理效率和灵活性。混合云作为一种新兴的云计算模式，将私有云和公有云结合起来，既能享受公有云的便利，又能保证私有数据的安全性，因此被越来越多的科学机构所采用。然而，在大规模数据处理中，数据安全问题一直是备受关注的焦点，科学工作流的正确执行依赖于数据的安全性。若数据丢失、泄露或被篡改，则会导致科学工作流中断、秘密信息泄露和错误的执行结果。在混合云环境下，数据存储和传输安全问题更为复杂。因此，如何在混合云环境下保证科学工作流的数据安全，成了一个亟待解决的问题。

针对隐私数据的保护，需要依靠安全技术和合理的管理策略。作为存储数据的重要场所之一，数据中心必须满足隐私数据的安全要求。然而，由于不同类型数据的安全需求不同，数据中心的安全级别应该根据其所能提供的安全服务进行分级，以为隐私数据提供更高的安全保障。目前，在数据中心建设方面，行业通常采用等级划分的方式来评估数据中心的整体性能。例如，美国 Uptime Institute 提出的等级分类系统将数据中心分为 Tier Ⅰ、Tier Ⅱ、Tier Ⅲ和 Tier Ⅳ四个等级，现已被广泛采用。在国内，GB 50174—2017 根据数据中心的使用性质、数据丢失或网络中断所造成的损坏和影响程度将数据中心定义为 A、B、C 级。

然而，传统的数据布局策略通常将隐私数据固定存储在单一数据中心中，这会导致一些安全问题，如单点故障等。不同的数据具有不同的安全需求，而且一些敏感的隐私数据必须存储在更加安全的数据中心中。因此，本节采用安全等级分级机制来确定数据需要的安全需求和数据中心所能提供的安全服务。这种方法可以确保隐私数据被妥善地保护，同时也能够使数据中心在保护隐私数据的同时，充分利用资源，提高数据传输效率。另外，当隐私数据集使用固定存储策略时，由于数据存储位置的限制，科学工作流任务只能从固定的数据中心中获取数据进行计算，而这些数据中心可能会因为网络拥堵或服务器负载高导致计算时间变长，从而影响整个科学工作流的执行时间。而采用安全等级分级机制存储方法可以将隐私数据存储在不同的安全等级数据中心中，这样一方面可以减轻单个数据中心的压力，另一方面可以在不同的数据中心中并行计算，从而提高科学工作流的执行效率。科学工作流执行过程中通常需要进行数据传输和处理，而这些数据通常存储在不同的数据中心中，其传输速度受到网络带宽、数据大小和传输协议等因素的影响。因此，通过减少数据传输时间可以大大缩短科学工作流的执行时间。

本节的主要贡献如下：

(1) 构建了一种混合云环境下面向数据安全的科学工作流数据布局模型，并分析了数据集的安全需求和数据中心所能提供的安全服务，提出了安全等级分级机制。

(2) 引入遗传算法的交叉算子和变异算子，在全局最优粒子替换中结合模拟退火算法的 Metropolitan 准则，有效避免粒子群算法陷入局部最优，提高了算法的收敛速度和全局最优解的精度。

(3) 考虑云间的带宽速度和数据中心的存储容量，设计了一种适用于本模型的基于 SAGA-PSO 的数据布局策略，从全局角度优化数据传输时延。

3.1.2　相关工作

近年来，随着云计算技术的普及和云计算强大的并行计算能力，许多科研工作者开始着手研究科学工作流在云计算环境下的数据布局问题，科学工作流数据布局策略的优劣直接影响到整个系统的作业性能。

在混合云环境下，科学工作流可能涉及多个数据中心或存储设备，有效管理数据传输对于整体性能至关重要。Deng 等提出了一种多级 K-cut 图分割算法，他将数据根据其依赖关系不断切割成子图，之后布局到对应数据中心中，在满足负载平衡和固定数据集约束的同时，最小化跨数据中心的数据传输总量。Liu 等考虑了数据的相关度，提出了基于相关度的两阶段数据放置策略。该策略在数据布局阶段将关系紧密的数据放在同一数据中心，将关系松散的数据放在不同数据中心，在任务执行时将任务调度到相关性最大的数据中心，从而减少了数据中心间的数据移动次数和数据传输量。Wang 等提出了一种基于动态计算相关性(DCC)的数据放置策略。该策略将具有高 DCC 的数据集放置在同一个数据中心，并在运行时动态地将新生成的数据集分布到最合适的数据中心，有效减少数据中心之间的数据调度数量。Kim 等则基于任务的依赖性提出了一种数据放置策略。在构建阶段，该策略将数据中心中的初始数据集进行分组，并在运行时将新生成的数据集布局到最合适的数据中心，可有效减少数据移动。Li 等研究者通过创建一个基于工作流的数据定位模型，引入了一个分为两步的数据分布方案。这个方案利用离散粒子群优化技术动态地将数据指派

给合适的数据中心，以此达到相比于传统的基于任务的数据定位策略更高的成本效益。Liu 等提出了一种智能数据放置策略，考虑数据中心的计算能力、存储预算、数据项相关性等因素将整个数据集划分为小的数据项，并在多个数据中心进行布局，在运行中生成中间数据时，采用线性判别分析将其放置在适当的数据中心上，最小化由数据移动引起的通信开销。

然而，现有研究聚焦于降低数据传输次数、数据传输时间和代价开销，在数据安全保护方面考虑不足，随着数据间交互越来越多，传统隐私数据进行单一固定的数据存储方案越发具有局限性，因此，需要采取更加有效的措施来保护隐私数据在科学工作流中的安全传输和存储。

3.1.3 问题模型

本小节具体介绍关于混合云环境下面向数据安全的科学工作流问题模型，其主要涉及科学工作流、混合云环境和数据布局器三种角色。图 3-1 展示了布局系统的框架图。

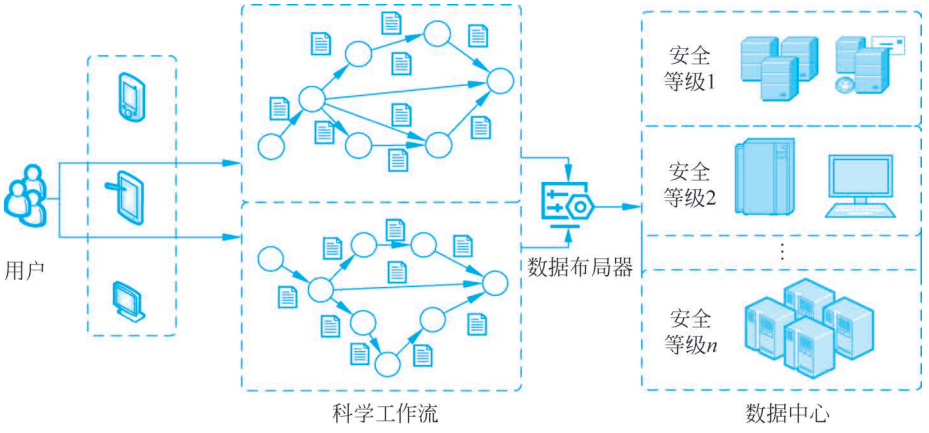


图 3-1 混合云环境下面向数据安全的科学工作流布局框架

1. 科学工作流

科学工作流可以用有向无环图 $G=(T,\boldsymbol{E},S)$ 表示，其中 $T=\{t_1,t_2,\cdots,t_n\}$ 为科学工作流中的任务集合；$\boldsymbol{E}=\{e_{12},e_{13},\cdots,e_{ij}\}$ 为任务之间的控制依赖矩阵，$e_{ij}=1$ 表示 t_i 是 t_j 的前驱节点，$e_{ij}=0$ 表示 t_j 是 t_i 的后继节点；$S=\{s_1,s_2,\cdots,s_n\}$ 为科学工作流中执行时所需的数据集合。

某个数据集 s_i 可表示为

$$s_i=\{z_i,o_i,r_i,w_i\} \tag{3-1}$$

式中，z_i 表示数据集 s_i 大小；o_i 表示数据集 s_i 的来源，若 s_i 为初始数据集 S_{ini}，o_i 记为 0，若 s_i 为生成数据集 S_{gen}，o_i 表示生成数据集 s_i 的任务；r_i 表示数据集 s_i 布局的数据中心位置；w_i 表示数据集 s_i 的安全等级。

对数据集进行安全等级分级，需要综合考虑数据的保密性、完整性和可用性等方面的需求。首先，根据数据集的敏感程度，将其分为非隐私数据和隐私数据两类，这里规定非隐私数据的安全等级最低，设为 0，而隐私数据具有不同的且高于非隐私数据的安全等级，如

1、2、3 等。数据集的安全等级越高，则说明该数据集的安全需求越高。

某个任务 t_i 可表示为

$$t_i=\{I_i,O_i,f(t_i),m_i\} \tag{3-2}$$

式中，I_i 为任务 t_i 的输入数据集合；O_i 为任务 t_i 的输出数据集合；$f(t_i)$表示执行任务 t_i 的数据中心位置；m_i 表示任务 t_i 的安全等级。

任务的安全等级可以根据其所涉及的数据集的安全等级来进行设置。当任务需要输入和输出数据集时，其安全等级应不低于其输入和输出数据集中最高的安全等级，以确保与其关联的数据安全性。因此，将任务的安全等级取值为 I_i 和 O_i 中最高的安全等级。

2. 混合云环境

混合云环境 C 中包含公有云 C_{pub} 和私有云 C_{pri}，二者均由多个数据中心组成，公有云 $C_{\text{pub}}=\{c_1,c_2,\cdots,c_n\}$ 包含 n 个数据中心，私有云 $C_{\text{pri}}=\{c_1,c_2,\cdots,c_m\}$ 包含 m 个数据中心。数据中心 c_i 可表示为

$$c_i=\{h_i,v_i,p_i,l_i\} \tag{3-3}$$

式中，h_i 表示数据中心的存储容量上限，公有云的容量不设上限，私有云上的数据集大小不得超过该存储容量；v_i 表示数据中心 c_i 当前已使用的存储容量，将已使用存储容量超出容量限制的数据中心称为超负荷数据中心；p_i 表示数据中心 c_i 存储单位容量所需的代价；l_i 表示该数据中心的安全等级。

数据中心的安全等级需要综合考虑数据中心的地理位置、安全措施、可靠性等方面的因素，不同云数据中心提供不同安全服务来保护存储在其上的数据集。根据数据中心提供的安全服务，将数据中心进行安全等级分级。由于用户与公有云服务商之间难以相互信任，且安全性问题一直是公有云端不可忽视的问题，而将数据存储在私有云上，用户对基础设施拥有更多控制权，可以通过局域网、防火墙等技术屏蔽外部访问。因此，定义公有云中数据中心安全等级最低，设为 0，私有云中数据中心的安全等级高于公有云，如 1、2、3 等。

3. 数据布局器

科学工作流在执行前，由数据布局器将数据集放置到合适的数据中心中，生成数据布局方案。为避免数据的泄露和被篡改等安全威胁，将数据的安全需求和数据中心提供的安全服务进行量化，根据所提供的安全等级分级机制进行匹配，从而提高隐私数据的安全性。

1）安全等级分级机制

安全等级分级机制如下：当数据集 s_i 和任务 t_j 放置或布局在数据中心 c_k 中，必须满足

$$w_i,m_j\leqslant l_k \tag{3-4}$$

即仅允许数据集和任务放置或布局在安全等级不低于它自身的数据中心中。同样地，数据中心仅允许存储或接受安全等级不高于它自身的数据集和任务。

综上所述，对于混合云环境 C 中的任务 t_j，任务 t_j 放置在数据中心 c_k。当 t_j 的输入数据集或输出数据集的集合$\{I_j,O_j\}$中包含数据集 s_i，且 $w_i=m_j=L(L\geqslant 1)$，$r_i=A(A\in[1,|C|])$，必须满足：

① $r_i\equiv A$；

② 对于数据中心 c_k，总有 $l_k\geqslant L$；

③ 对于数据中心 c_k 及任意的数据集 $s_n \in \{I_j, O_j\}$，总有 $l_k \geqslant w_n$。

对上述规则做如下解释：对①，数据集存储在数据中心后，在后续任务进行中其存储位置不可改变；对②，此时任务的安全等级为 L，放置任务的数据中心的安全等级不低于任务的安全等级；对③，由于 $l_k \geqslant L$，且 $w_i = L \geqslant w_n$，所以 $l_k \geqslant w_n$。

2）数据布局方案

在本节中主要关注数据集跨数据中心产生的传输时延，科学工作流执行过程中数据存储、读取数据等时延与上述时延相比可以忽略不计。数据集 s_k 布局到数据中心 c_i 的映射 y 可表示为 $y=(s_k, c_i)$，我们使用 Y 来表示所有数据集到数据中心的映射集合。数据集 s_k 从数据中心 c_i 传输到数据中心 c_j 的单次数据传输可以表示为 $\mathrm{tra}_n=(c_i, s_k, c_j)$，那么单次跨数据中心传输产生的传输时间可表示为：

$$q(\mathrm{tra}_n)=q(c_i, s_k, c_j)=\frac{z_k}{b_{ij}} \tag{3-5}$$

式中，b_{ij} 为数据中心 c_i 和数据中心 c_j 之间的带宽，$b_{ij}=b_{ji}$。$\mathrm{Tras}=\{\mathrm{tra}_1, \mathrm{tra}_2, \cdots, \mathrm{tra}_n\}$，为所有数据集的跨数据中心传输集合。因此，$q(\mathrm{tra}_n)=q(c_i, s_k, c_j)=\frac{z_k}{b_{ij}}$ 总传输时间 Q 可表示为：

$$Q=\sum_{n=1}^{|\mathrm{Tras}|} q(\mathrm{tra}_n) \tag{3-6}$$

综上所述，整个数据布局方案可定义为

$$N=(S, C, Y, Q) \tag{3-7}$$

基于上述定义，混合云环境下面向数据安全的科学工作流数据布局方法可表示为式(3-8)和式(3-9)，其核心目标为降低传输时延，即总传输时间 Q 最低，且同时满足以下两个条件约束：①数据中心的存储容量约束；②安全等级约束。在下文中这两个约束条件合称为“条件约束”。

$$\min Q \tag{3-8}$$

$$\begin{cases} \text{s.t. } \forall j, & \sum_{i=1}^{|S|} z_i \cdot u_{ij} \leqslant h_j \\ \text{s.t. } \forall i, j, & w_i \cdot u_{ij} \leqslant l_j \end{cases} \tag{3-9}$$

式中，$u_{ij}=\{0,1\}$，表示数据集 s_j 是否存储在数据中心 c_i 上，如果是，则 $u_{ij}=1$，否则 $u_{ij}=0$。

3.1.4 基于 SAGA-PSO 的数据布局策略

下面首先介绍本节所使用的问题编码，然后给出算法的适应度函数，最后提出基于遗传算法和模拟退火算法的自适应粒子群优化算法(SAGA-PSO)的科学工作流数据布局策略的具体描述，在满足条件约束的情况下，尽可能减少科学工作流的传输时间。

1. 问题编码

本小节采用离散型编码方式对数据的布局策略进行编码，每个粒子代表一种布局策略，粒子 i 在 t 次迭代的位置为

$$X_{ik}^{t}=(x_{i1}^{t},x_{i2}^{t},\cdots,x_{in}^{t}) \tag{3-10}$$

每个粒子由 n 个分位组成，n 代表数据集数量。$X_{ik}^{t}(k=1,2,\cdots,n)$表示第 i 个粒子在第 t 次迭代时第 k 个数据的存储位置，取值范围为$[1,|C|]$。

$$X_{2}^{7}=(2,1,3,2,4,2,1,2) \tag{3-11}$$

式(3-11)为一个粒子的编码示例，该粒子的数据集数量 n 为 8。图 3-2 表示该编码粒子对应的数据布局位置。从中可知：编号 4 的数据放置在服务器 2 上。

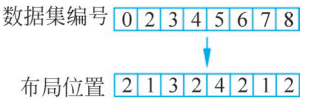


图 3-2　问题编码

2. 适应度函数

本节旨在优化数据传输时延，粒子对应数据布局结果的传输时间越短，该粒子质量越高，但是在编码过程中可能存在某些不可行解粒子，不可行解的原因是该粒子不满足条件约束。将导致粒子不可行的数据合称为非法数据集 $S_{\inf}$，因此对于可行解粒子和不可行解粒子的优劣比较主要分以下三种情况：

(1) 两个粒子皆为可行解粒子，优先选择传输时间较短的粒子。粒子的适应度函数可定义为

$$\text{fitness}=Q \tag{3-12}$$

(2) 两个粒子皆为不可行解粒子，同样选择传输时间低的粒子，因为该粒子在后续更新操作中更可能变为可行解粒子。适应度函数与式(3-12)一致。

(3) 两个粒子一个为可行解粒子，一个为不可行解粒子，选择可行解粒子。粒子适应度函数可定义为

$$\text{fitness}=\begin{cases}0, & \text{满足条件约束的粒子}\\ 1, & \text{其他}\end{cases} \tag{3-13}$$

3. 粒子的更新策略

PSO 算法通过模拟鸟群在捕食区域内捕食唯一食物，建立粒子群模型，之后迭代收敛求最优解。在传统 PSO 算法中，Lin 等结合遗传算法的变异算子和交叉算子思想进行改进(GA-PSO)，为使该算法适应本模型，粒子 i 在第 t 次迭代进行如下更新：定义变异操作的运算符⊕如式(3-14)所示，定义交叉操作的运算符⊗如式(3-15)和式(3-16)所示。

$$A_{i}^{t}=\omega\oplus M(X_{i}^{t-1})=\begin{cases}M(X_{i}^{t-1}), & r_{0}<\omega\\ X_{i}^{t-1}, & \text{其他}\end{cases} \tag{3-14}$$

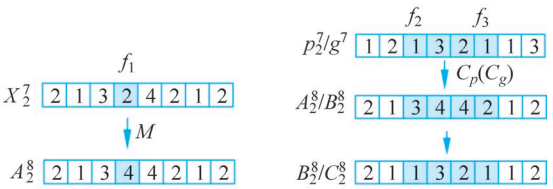
$$B_{i}^{t}=c_{1}\otimes C_{p}(A_{i}^{t},p_{i}^{t-1})=\begin{cases}C_{p}(A_{i}^{t},p_{i}^{t-1}), & r_{1}<c_{1}\\ A_{i}^{t}, & \text{其他}\end{cases} \tag{3-15}$$

$$C_{i}^{t}=c_{2}\otimes C_{g}(B_{i}^{t},g^{t-1})=\begin{cases}C_{g}(B_{i}^{t},g^{t-1}), & r_{2}<c_{2}\\ B_{i}^{t}, & \text{其他}\end{cases} \tag{3-16}$$

式中，$r_{0},r_{1},r_{2}\in(0,1)$，为随机因子；$M()$为变异操作，随机选取编码粒子中的某个分位，改变该分位的数值；p_{i} 和 g 分别为个体最优粒子和全局最优粒子；$C_{p}()$和 $C_{g}()$分别表示与个体最优粒子和全局最优粒子的交叉过程，随机选取 A_{i}^{t} 和 B_{i}^{t} 的起止位置，分别与 p_{i}^{t-1}

和 g^{t-1} 相同分位之间的数值进行交叉。

图 3-3(a)表示粒子的变异操作，其中选取粒子的第 4 分位即 f_1 位置进行变异，其存储的数据中心编号从 2 变异为 4；图 3-3(b)表示粒子的交叉操作，该实例从第 3 分位(即 f_2)至第 6 分位(即 f_3)选择交叉，粒子的值从[3,4,4,2]变为[1,3,2,1]。



(a) 粒子的变异操作　　(b) 粒子的交叉操作

图 3-3　编码粒子的更新操作

需要注意的是，在粒子的初始化和变异过程中，数据布局后的数据集存储位置应满足安全等级约束。当粒子为不可行解粒子，优先选取存储在超负荷数据中心中的数据集分位进行变异，使其更有可能变异为可行解粒子。

综合上述分析，对粒子 i 在 t 时刻的更新操作可定义为

$$X_i^t = c_2 \oplus c_g(c_1 \oplus C_p(\omega \oplus M(X_i^{t-1}), p_i^{t-1}), g^{t-1}) \tag{3-17}$$

GA-PSO 虽在一定程度上提高了种群进化的多样性，但当陷入局部极值时，仍采取原 g^t 的更新策略，跳出局部极值的效率并不高，在迭代后期 g^t 更是难以进行有效更新。因此，结合模拟退火算法概率接受更差解的思想对 GA-PSO 进行改进，使其有效跳出局部最优，增大搜寻到更优解的概率。

在模拟退火算法(SA)中，使用 Metropolitan 准则以 Me 的概率接受比当前解更差的解。SA 的接受概率 Me 的大小为

$$Me = \mathrm{e}^{\frac{f(x_n)-f(x_0)}{T}} \tag{3-18}$$

式中，T 为当前模拟退火温度；$f(x_n)$ 和 $f(x_0)$ 分别为新解和旧解的模拟退火算法适应度值。

在每次迭代过程中，对全局最优粒子 g^t 进行如下更新操作：所有个体最优粒子 p_i^t 都将以 $f_{TF}(p_i^t)$ 的概率替换全局最优粒子 g^t。$f_{TF}(p_i^t)$ 为模拟退火算法适应度值，其大小定义为

$$f_{TF}(p_i^t) = \frac{\mathrm{e}^{-\frac{f(p_i^t)-f(g^t)}{T}}}{\sum_{i=1}^{n} \mathrm{e}^{-\frac{f(p_i^t)-f(g^t)}{T}}} \tag{3-19}$$

式中，$f(p_i^t)$ 和 $f(g^t)$ 为个体最优粒子和全局最优粒子的粒子群优化算法适应度值。

在获得 p_i^t 模拟退火算法适应度值后使用轮盘赌从所有 p_i^t 中选择替换 g^t。在很大程度上避免了陷入局部最优，且在粒子向种群学习的过程中，扩大了解的搜索范围，在算法快速收敛的同时仍保持较高的种群多样性。

4. 粒子到数据布局方案的映射

算法 3.1 描述了粒子到数据布局方案的映射过程。输入为混合云环境 C，科学工作流

$G=\{T,E,S\}$和编码粒子 X，输出为数据布局方案 $N=\{S,C,Y,Q\}$。需要注意的是，在任务执行过程中，若某中间数据集后续不再被使用，那么该数据集可以被删除，以增大数据中心可使用容量，节约空间。

算法 3.1　编码粒子到数据布局方案的映射

```
输入：G,C,X
输出：N,Q
0：  初始化：Q←0，∀a∈[1,|C|]，v_a←0
1：  for each S_i of S_ini do
2：      根据初始数据集 S_ini 更新数据中心已使用容量
3：      if 当前数据中心超负荷 then
4：          将当前粒子设为不可解粒子，更新 S_ini
5：      end if
6：  end for
7：  for j←1 to j=|T| do
8：      任务 t_j 调度到安全等级约束下传输时间最小的数据中心 c_k
9：      IO_i=I_i+O_i
10：     if IO_i+v_k>h_k then
11：         将当前粒子设为不可解粒子，更新 D_inf
12：     end if
13：     将任务 t_j 的输出数据集传输到对应数据中心 c_m
14：     更新数据中心 c_m 已使用容量 v_m
15：     for j←1 to j=|IO_i|
16：         根据式(3-5)计算数据集 s_j 在对应数据中心 c_a 和 c_b 的数据传输时间 q(c_a,s_j,c_b)
17：         Q+=q(c_a,s_j,c_b)
18：     end for
19：     for each S_l of S_gen
20：         if S_l 后续不再被使用 then
21：             delete S_l
22：         更新存储 S_l 的数据中心已使用容量
23：     end for
24： end for
25： return N,Q
26： end procedure
```

算法详细描述如下：将所有数据中心的存储量初始化为 0，总传输时间 Q 初始为 0(第 0 行)。遍历初始数据集(第 2～7 行)，将初始数据集布局到其对应的数据中心，并更新该数据中心的存储量，若当前存储量超出数据中心的容量限制，则该编码粒子为不可行解粒子，将存放在该数据中心的数据集存入 D_{inf}，方便后续变异。按任务执行顺序遍历任务(第 8～25 行)，在任务执行前，根据贪心策略，将任务调度到安全等级约束下传输时间花费最小的数据中心(第 9 行)，并判断该数据中心当前的存储量和该任务的输入/输出数据集大小之和是否超过该数据中心的容量 h_k(第 10～13 行)，若超过，则数据中心容量不足，该编码粒子为不可解粒子，将存放在该数据中心的数据集存入 D_{inf}；否则任务正常执行。将输出数据集进行输出，更新数据中心已使用容量(第 14～15 行)。遍历当前任务的输入/输出数据集

IO_i(第 16～19 行),根据式(3-6)计算当前数据集 s_j 跨数据中心 c_a 和数据中心 c_b 的传输时间 $q(c_a,s_j,c_b)$,累加即得总传输时间 Q。遍历所有中间数据集(第 20～24 行),如果该中间数据集 S_l 在后续任务中不再被使用,则删除 S_l,并更新对应数据中心的已使用容量,输出数据布局方案 N 和总传输时间 Q(第 34 行)。

5. 参数设置

惯性权重因子 ω 决定 PSO 算法的速度的变化情况。通过自适应调整机制,根据当前粒子优劣性即当前粒子与全局最优粒子之间的差异程度对 ω 进行自适应调整。ω 可表示为

$$\omega=\omega_{\max}-(\omega_{\max}-\omega_{\min})\times\exp\left(\frac{d(X^{t-1},g^{t-1})}{|S|}\Big/\frac{d(X^{t-1},g^{t-1})}{|S|}-1.01\right) \tag{3-20}$$

式中,$d(X^{t-1},g^{t-1})$表示 X^{t-1} 与 g^{t-1} 在相同分位上取值不同的个数。

自身认知因子 c_1 和种群认知因子 c_2 分别设为

$$c_1=c_1^s-(c_1^s-c_1^e)\times(i_c/i_m) \tag{3-21}$$

$$c_2=c_2^s-(c_2^s-c_2^e)\times(i_c/i_m) \tag{3-22}$$

式中,c_1^s 和 c_1^e、c_2^s 和 c_2^e 分别为自身认知因子 c_1 和种群认知因子 c_2 的设定初始值与最终值;i_c 和 i_m 为当前迭代次数和最大迭代次数。

退火操作降温公式为

$$T=\frac{T_0}{1+t} \tag{3-23}$$

式中,T_0 为初始温度;t 为迭代次数。

6. 算法流程

基于前面所述,本节所提出的 SAGA-PSO 算法的具体流程如图 3-4 所示。

3.1.5 实验

实验结果的讨论将在这一节进行,主要围绕以下研究问题(Research Question,RQ)进行讨论。

RQ1:在条件约束下,和其他布局算法相比,SAGA-PSO 在降低科学工作流的传输时延方面是否具有优越性。

RQ2:在相同实验环境下,仅改变私有云数据中心容量,观察分析不同算法在不同科学工作流执行后的实验结果。

RQ3:在一致的实验设置中,通过仅调整私有云数据中心的数量,进行了对比分析,以观察不同算法在执行各种科学工作流之后的实验成效。

RQ4:在相同实验环境下,仅改变数据中心间的带宽,观察分析不同算法在代表性科学工作流执行后的实验结果。

RQ5:为探讨本节所提供的安全等级分级的重要性,分析比较隐私数据集固定存储与安全等级分级存储的不同算法在不同科学工作流执行后的实验结果。

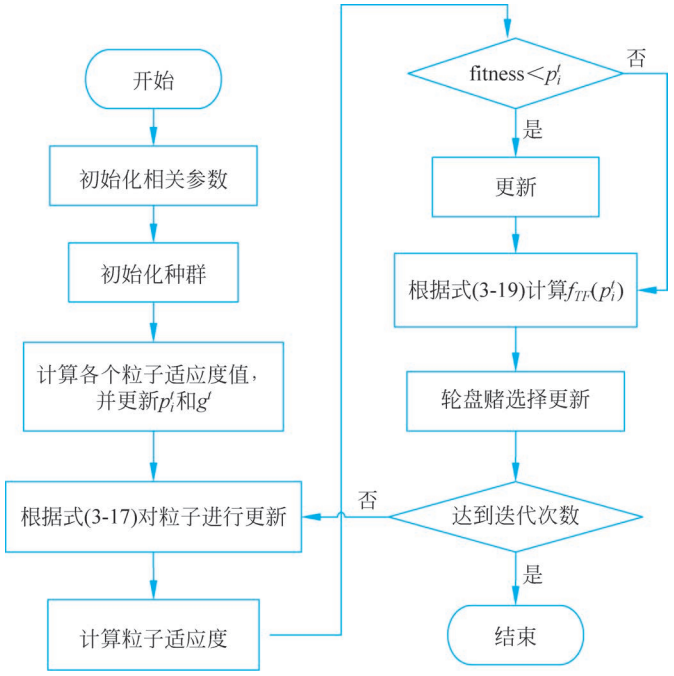


图 3-4　SAGA-PSO 算法流程图

1. 实验设置

本节实验都在 16GB RAM 和 2.6GHz Intel i7-6700HQ CPU 的 Windows 10 系统下运行，SAGA-PSO 和其他对比算法均在 Python 3.8 环境下实现。

本实验涉及 SAGA-PSO 的相关参数如表 3-1 所示。科学工作流主要来源于 3 个不同科学领域的中型科学工作流模型：Montage、Inspiral、Epigenomics。规模大小为中型科学工作流（约 50 个任务数）和大型科学工作流（约 100 个任务数），它们拥有其独特的任务依赖结构和计算需求。

表 3-1　SAGA-PSO 的相关参数

参　　数	值
种群大小	50
最大迭代次数	1000
$\omega_{\max}$，$\omega_{\min}$	0.9，0.4
c_1^s，c_1^e	0.9，0.2
c_2^s，c_2^e	0.9，0.4
T_0	1000

特别地，为更好地观察不同影响因素对实验的影响，设置默认实验环境如下：混合云环境中有 4 个数据中心$\{c_0, c_1, c_2, c_3\}$，其中 c_0 为容量无限，安全等级为 0 的公有云数据中心；c_1、c_2、c_3 分别为安全等级为 1、2、3 的私有云数据中心。定义私有云数据中心基准容量为

$$h = \delta \cdot \frac{\sum_{i=1}^{|S|} Z_i}{|C|-1} \tag{3-24}$$