(19) Deutsches Patent- und Markenamt

(10) **DE 11 2018 002 920 T5** 2020.04.09

(12) # Veröffentlichung

der internationalen Anmeldung mit der
(87) Veröffentlichungs-Nr.: **WO 2018/226637**
in der deutschen Übersetzung (Art. III § 8 Abs. 2 IntPatÜG)
(21) Deutsches Aktenzeichen: **11 2018 002 920.4**
(86) PCT-Aktenzeichen: **PCT/US2018/035976**
(86) PCT-Anmeldetag: **05.06.2018**
(87) PCT-Veröffentlichungstag: **13.12.2018**
(43) Veröffentlichungstag der PCT Anmeldung in deutscher Übersetzung: **09.04.2020**

(51) Int Cl.: ***G06F 8/34*** (2018.01)
G06F 16/00 (2019.01)

(30) Unionspriorität:

62/516,396	**07.06.2017**	**US**
15/997,278	**04.06.2018**	**US**

(71) Anmelder:
Ab Initio Technology LLC, Lexington, Mass., US

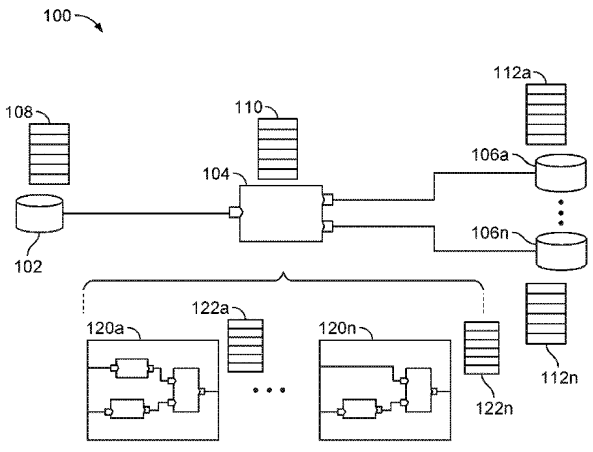
(74) Vertreter:
BOEHMERT & BOEHMERT Anwaltspartnerschaft mbB - Patentanwälte Rechtsanwälte, 28209 Bremen, DE

(72) Erfinder:
Epstein, Richard A., Yarmouth, ME, US; Palmer, Mike, Raleigh, NC, US

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **DATENFLUSSGRAPHENKONFIGURATION**

(57) Zusammenfassung: Ein Verfahren zum Konfigurieren eines ersten computerausführbaren Programms beinhaltet Folgendes: Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche; Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet. Das Verfahren beinhaltet automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt. Das Verfahren beinhaltet automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. Das Konfigurieren beinhaltet Folgendes: Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.



Beschreibung

Prioritätsanspruch

[0001] Die vorliegende Anmeldung beansprucht die Priorität der US-Patentanmeldung mit der Serien-Nr. 62/516,396, eingereicht am 7. Juni 2017, und der US-Patentanmeldung mit der Serien-Nr. 15/997,278, eingereicht am 4. Juni 2018, deren Inhalte hiermit unter Bezugnahme in ihrer Gesamtheit aufgenommen werden.

HINTERGRUND

[0002] Eine Benutzeroberfläche kann erzeugt und einem Benutzer angezeigt werden, um zu ermöglichen, dass der Benutzer mit Daten interagiert, die durch ein Rechensystem verarbeitet werden. Manche Benutzeroberflächen sind statische Komponenten eines Anwendungsprogramms und die Benutzeroberflächen arbeiten auf die gleiche Weise für mehrere Benutzer des Problems. Manche Benutzeroberflächen können durch einen Benutzer des Anwendungsprogramms definiert werden, so dass die Benutzeroberfläche für einen bestimmten Zweck maßgeschneidert ist. Zum Beispiel kann ein Anwendungsprogramm die Spezifizierung mehrerer Benutzeroberflächen ermöglichen und kann ein Benutzer zwischen den mehreren Benutzeroberflächen wählen.

KURZBESCHREIBUNG

[0003] Bei einem Aspekt beinhaltet ein Verfahren zum Konfigurieren eines ersten computerausführbaren Programms Folgendes: Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche; Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet. Das Verfahren beinhaltet automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt. Das Verfahren beinhaltet automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. Das Konfigurieren beinhaltet Folgendes: Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

[0004] Ausführungsformen können eines oder mehrere der folgenden Merkmale beinhalten.

[0005] Das erste computerausführbare Programm beinhaltet einen ersten Datenflussgraphen, der einen oder mehrere Knoten, die Datenverarbeitungskomponenten repräsentieren, und eine oder mehrere Verknüpfungen, die Flüsse von Daten zwischen den Datenverarbeitungskomponenten repräsentieren, beinhaltet.

[0006] Das zweite computerausführbare Programm beinhaltet einen zweiten Datenflussgraphen, der einen oder mehrere Knoten, die Datenverarbeitungskomponenten repräsentieren, und eine oder mehrere Verknüpfungen, die Flüsse von Daten zwischen den Datenverarbeitungskomponenten repräsentieren, beinhaltet.

[0007] Die Informationen, die die Quelle von Daten angeben, identifizieren einen Typ der Quelle von Daten. Das Empfangen der Informationen, die die Quelle von Daten angeben, beinhaltet Empfangen einer Auswahl des Typs der Quelle von Daten aus einem Satz von einem oder mehreren Typen. Der Satz von einem oder mehreren Typen beinhaltet eine Datei, eine Tabelle, einen HDFS-Datensatz, eine HIVE-Tabelle und/oder eine Warteschlange.

[0008] Die Informationen, die die Quelle von Daten angeben, beinhalten Informationen, die einen Ort der Quelle von Daten angeben.

[0009] Die Informationen, die das Datenziel angeben, identifizieren einen Typ des Datenziels. Das Empfangen der Informationen, die das Datenziel angeben, beinhaltet Empfangen einer Auswahl des Typs des Datenziels aus einem Satz von einem oder mehreren Typen. Der Satz von einem oder mehreren Typen beinhaltet eine Datei, eine Tabelle, einen HDFS-Datensatz, eine HIVE-Tabelle und/oder eine Warteschlange.

[0010] Die Informationen, die das Datenziel angeben, beinhalten Informationen, die mehrere Datenziele angeben.

[0011] Die Informationen, die das Datenziel angeben, beinhalten Informationen, die einen Ort des Datenziels angeben.

[0012] Die Charakterisierung des Prozesses beinhaltet eine Charakterisierung von jedem von mehreren Prozessen.

[0013] Das Verfahren beinhaltet Bestimmen der Charakteristiken, die mit dem Prozess assoziiert sind, basierend auf dem Typ des Prozesses.

[0014] Das Empfangen einer Charakterisierung eines Prozesses beinhaltet Empfangen einer Auswahl eines speziellen Prozesses aus einem Satz von einem oder mehreren Prozessen. Die Prozesse des Satzes beinhalten einen Zuordnungsprozess, einen Filterprozess, einen Deduplizierungsprozess, einen Aggregationsprozess, einen Replikationsprozess, einen Normalisierungsprozess, einen Unterteilungs- und Sortierungsprozess, einen Datenerfassungsänderungsprozess und/oder einen Zusammenführungsprozess.

[0015] Das Verfahren beinhaltet automatisches Konfigurieren des ersten Datenflussgraphen dazu, den zweiten Datenflussgraphen zu einer Laufzeit des ersten Datenflussgraphen zu referenzieren.

[0016] Das automatische Konfigurieren des ersten Datenflussgraphen dazu, den zweiten Datenflussgraphen zu referenzieren, beinhaltet Konfigurieren des ersten Datenflussgraphen dazu, den zweiten Datenflussgraphen während einer Ausführung des ersten Datenflussgraphen auszuführen.

[0017] Bei einem Aspekt speichert ein nichtflüchtiges computerlesbares Medium Anweisungen zum Bewirken, dass ein Rechensystem ein erstes computerausführbares Programm konfiguriert. Die Anweisungen beinhalten Anweisungen, um das Rechensystem zu Folgendem zu veranlassen: Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche; Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet; automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. Das Konfigurieren beinhaltet Folgendes: Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

[0018] Bei einem Aspekt beinhaltet ein Rechensystem zum Konfigurieren eines ersten computerausführbaren Programms einen oder mehrere Prozessoren und einen Speicher, wobei der eine oder die mehreren Prozessoren und der Speicher zu Folgendem konfiguriert sind: Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche; Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet; automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. Das Konfigurieren beinhaltet Folgendes: Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

[0019] Bei einem Aspekt beinhaltet ein Rechensystem zum Konfigurieren eines ersten computerausführbaren Programms Folgendes: ein Mittel zum Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche; ein Mittel zum Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet; ein Mittel zum automatischen Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und ein Mittel zum automatischen Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. Das Konfigurieren beinhaltet Folgendes: Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

Figurenliste

Fig. 1 ist ein Systemdiagramm.

Fig. 2 ist ein Diagramm eines Datenflussgraphen.

Fig. 3 ist ein Diagramm einer Benutzeroberfläche.

Fig. 4 ist ein Flussdiagramm.

Fig. 5A-5G sind Beispiel-Screenshots.

Fig. 6-8 sind Beispiel-Screenshots.

Fig. 9A und **Fig. 9B** sind Beispiel-Screenshots.

Fig. 10 ist ein Systemdiagramm.

Ausführliche Beschreibung

[0020] Wir beschreiben hier einen Ansatz zum Konfigurieren eines Datenflussgraphen basierend auf Charakteristiken, die durch einen Benutzer bereitgestellt werden. Die Charakteristiken können eine Quelle von Daten, die durch den Datenflussgraphen zu verarbeiten sind, ein Ziel für Daten, die von dem Datenflussgraphen ausgegeben werden, definieren und können die Verarbeitung definieren, die an den Daten durch den Datenflussgraphen durchzuführen ist. Der Typ einer durchzuführenden Verarbeitung kann aus einer Bibliothek verfügbarer Optionen ausgewählt werden. Ein allgemeiner wiederverwendbarer Datenflussgraph wird gemäß diesen Charakteristiken parametrisiert.

[0021] Unter Bezugnahme auf **Fig. 1** beinhaltet ein Beispielsystem **10** zum Konfigurieren von Datenflussgraphen eine Datenquelle **12**, die eine oder mehrere Quellen von Daten, wie etwa Speicherungsvorrichtungen oder Verbindungen zu Online-Datenströmen, beinhalten kann, die jeweils Daten in einem beliebigen einer Vielzahl von Speicherungsformaten (z. B. Datenbanktabellen, Spreadsheet(Tabellenkalkulation)-Dateien, Flat-Textdateien oder einem nativen Format, das durch einen Mainframe verwendet wird) speichern können. Eine Ausführungsumgebung **14** beinhaltet ein Graphenkonfigurationsmodul **16** und ein Benutzeroberflächenmodul **22**. Die Ausführungsumgebung **14** kann auf einem oder mehreren Mehrzweckcomputern unter der Steuerung eines geeigneten Betriebssystems, wie etwa des UNIX-Betriebssystems, gehostet werden. Die Ausführungsumgebung **14** kann zum Beispiel eine Mehrknoten-Parallelrechenumgebung beinhalten, einschließlich einer Konfiguration von Computersystemen, die mehrere Zentralverarbeitungseinheiten (CPUs) verwenden, entweder lokal (z. B. Mehrprozessorsysteme, wie etwa SMP-Computer) oder lokal verteilt (z. B. mehrere Prozessoren, die als Cluster oder MPPs gekoppelt sind) oder entfernt oder entfernt verteilt (z. B. mehrere Prozessoren, die über LAN- oder WAN-Netze gekoppelt sind) oder eine beliebige Kombination davon.

[0022] Das Graphenkonfigurationsmodul **16** ändert die Konfiguration von Datenflussgraphen, wie unten ausführlicher beschrieben ist. Das Benutzeroberflächenmodul **22** zeigt einem Benutzer **30** Konfigurationsinformationen an und empfängt Konfigurationshandlungen von dem Benutzer **30**. Das Benutzeroberflächenmodul **22** kommuniziert auch mit dem Graphenkonfigurationsmodul **16**, das Datenflussgraphen basierend auf den Handlungen des Benutzers konfiguriert. Zum Beispiel können die Datenflussgraphen in der Datenquelle **12** gespeichert werden. Speichervorrichtungen, die die Datenquelle **12** bereitstellen, können lokal zu der Ausführungsumgebung **14** sein, zum Beispiel auf einem Speicherungsmedium gespeichert sein, das mit einem Computer, der die Ausführungsumgebung **14** ausführt, verbunden ist (z. B. einer Festplatte **18**), oder können entfernt zu der Ausführungsumgebung **14** sein, zum Beispiel auf einem entfernten System (z. B. einem Mainframe **20**), das sich über ein Local- oder Wide-Area-Datennetz in Kommunikation mit einem Computer befindet, gehostet werden, der die Ausführungsumgebung **14** ausführt.

[0023] Die Ausführungsumgebung befindet sich in Kommunikation mit einem Datenspeicherungssystem **26**, das Informationen enthält, die durch das Benutzeroberflächenmodul **22** verwendet werden, um eine Benutzeroberfläche anzuzeigen. Das Datenspeicherungssystem **26** ist auch für eine Entwicklungsumgebung **28** zugänglich, in der ein Entwickler **30** dazu in der Lage ist, Benutzeroberflächen zu entwickeln, die in dem Datenspeicherungssystem **26** gespeichert werden, das von dem Benutzeroberflächenmodul **22** verwendet wird, um eine Benutzeroberfläche anzuzeigen.

[0024] Die Datenquelle **12** ist bei manchen Implementierungen ein System zur Entwicklung von Anwendungen als Datenflussgraphen, die Ecken (Komponenten oder Datensätze) beinhalten, die durch gerichtete Verknüpfungen (die Flüsse von Arbeitselementen repräsentieren) zwischen den Ecken verbunden sind. Zum Beispiel ist eine derartige Umgebung ausführlicher in der US-Veröffentlichung mit der Nr. 2007/0011668, mit dem Titel „Managing Parameters for Graph-Based Applications“ beschrieben, die hiermit durch Bezugnahme aufgenommen wird.

[0025] Ein Datenflussgraph ist ein Computerprogramm, das innerhalb einer Datenflussgraphenausführungsumgebung ausgeführt wird, die Daten von einer oder mehreren Datenquellen verarbeitet. Die Daten von den Datenquellen werden gemäß dem Datenflussgraphen manipuliert und verarbeitet und an eine oder mehrere Datensenken exportiert. Datenquellen und -senken können zum Beispiel Dateien,

Datenbanken, Datenströme oder Warteschlangen beinhalten. Datenflussgraphen werden als gerichtete Graphen repräsentiert, die Knoten, die Datenverarbeitungskomponenten repräsentieren, die jeweils Code zum Verarbeiten von Daten von wenigstens einer Dateneingabe beinhalten und Daten an wenigstens eine Datenausgabe liefern, und Knoten, die Datensatzobjekte zum Zugreifen auf die Datenquellen und/oder -senken repräsentieren, beinhalten. Die Knoten sind durch gerichtete Verknüpfungen repräsentiert, die Flüsse von Daten zwischen den Komponenten repräsentieren, die bei den Datenquellen entspringen und bei den Datensenken enden. Die Datenausgabeports von vorgelagerten Komponenten sind mit den Dateneingabeports von nachgelagerten Komponenten verbunden. Die Datenflussgraphen können für unterschiedliche Datenquellen und unterschiedliche Datensenken wiederverwendet werden, die durch die Datensatzobjekte repräsentiert werden. Die Datenstrukturen und der Programmcode, die/der zum Implementieren von Datenflussgraphen verwendet werden/wird, kann/können mehrere unterschiedliche Konfigurationen unterstützen, indem sie/er parametrisiert wird/werden, um zu ermöglichen, dass zum Beispiel unterschiedliche Quellen und Senken einfach substituiert werden. Des Weiteren kann bei manchen Ausführungsformen der Fluss des Datenflussgraphen durch die Verwendung von Parametern verändert werden, so dass eine Komponente oder eine Reihe von Komponenten umgangen werden kann. Allgemein repräsentiert ein Parameter eine Eigenschaft eines Datenflussgraphen, die konfiguriert oder geändert werden kann. Zum Beispiel kann eine Eigenschaft zwischen Verwendungen des Datenflussgraphen geändert werden und kann der Datenflussgraph Operationen durchführen, die als Ergebnis der Änderung unterschiedlich sind. Ein Datenflussgraph kann zu einer Kompilierungszeit durch Verändern der Datenflussgraphenvorkompilierung, um eine spezielle Handlung durchzuführen, oder zu einer Laufzeit durch Einstellen von Parametern oder Verändern von Konfigurationsdateien, die durch den Datenflussgraphen verwendet werden, konfiguriert werden.

[0026] Zum Beispiel unter Bezugnahme auf **Fig. 2** beinhaltet ein allgemeiner Datenflussgraph **100** eine Datenquelle **102**, eine Komponente **104** und Datensenken **106-a-106n**. Die Datenquelle **102** ist mit dem Eingabeport der Komponente **104** verbunden. Der Ausgabeport der Komponente **104** ist mit den Datensenken **106a-106n** verbunden. Die Verbindungen zwischen der Datenquelle, der Komponente und den Datensenken definieren den Datenfluss durch den Datenflussgraphen.

[0027] Jede der Datenquelle **102**, der Komponente **104** und der Datensenken **106a-106n** ist jeweils mit einem Parametersatz **108**, **110**, **112a-112n** assoziiert, wobei jeder Parametersatz einen oder mehrere Parameter beinhaltet. Die Parameter definieren das Verhalten des Datenflussgraphen. Ein Parameter kann zum Beispiel den Ort der Datenquelle oder Datensenke auf einer physischen Platte definieren. Ein Parameter kann auch das Verhalten einer Komponente definieren, wie etwa wie eine Sortierkomponente die Dateneingabe in die Komponente sortiert. Bei manchen Beispielen werden Werte für die Parameter in einem Parametersatz bei einer Laufzeit des Datenflussgraphen besetzt.

[0028] Bei manchen Beispielen kann der Wert eines Parameters von dem Wert eines anderen Parameters abhängen. Beispielsweise kann eine Datenquelle in einer Datei in einem speziellen Verzeichnis gespeichert werden. Die für die Datenquelle eingestellten Parameter können einen ersten Parameter, der „VERZEICHNIS“ genannt wird, und einen zweiten Parameter, der „DATEINAME“ genannt wird, beinhalten. Bei diesem Beispiel hängt der DATEINAME-Parameter von dem VERZEICHNIS-Parameter ab (z. B. kann VERZEICHNIS „/usr/local“ sein und kann der DATEINAME „input.dat“ sein). Parameter können auch von den Parametern für andere Komponenten abhängen. Zum Beispiel kann die physische Position einer Datensenke für einen Datenflussgraphen von der physischen Position der Datenquelle für den Datenflussgraphen abhängen. Beispielsweise kann eine Datensenke einen Satz von Parametern beinhalten, der einen DATEINAME-Parameter beinhaltet, der von dem VERZEICHNIS-Parameter der Datenquelle abhängt (z. B. kann der DATEINAME-Parameter für die Datensenke „/usr/local/output.dat“ sein, wobei der Wert „/usr/local/“ aus dem VERZEICHNIS-Parameter für die Datenquelle erhalten wird).

[0029] Die Komponente **104** kann eine Graphenoberflächenkomponente sein, die einen oder mehrere andere Datenflussgraphen **120a-120n**, die manchmal als Subgraphen bezeichnet werden, referenziert. Zur Laufzeit lädt der Datenflussgraph **100** den (die) Subgraphen, die durch die Komponente **104** referenziert werden, dynamisch und führt diesen (die) aus, wodurch z. B. ermöglicht wird, dass der Datenflussgraph **100** flexibel auf verschiedene Funktionalitäten zugreift, die durch die Subgraphen **120a-120n** bereitgestellt werden. Ein oder mehrere Parameter der Komponente **104** definieren den (die) speziellen Subgraphen **120a-120n**, die durch die Komponente **104** referenziert sind. Jeder Subgraph **120a-120n** ist auch mit einem Parametersatz **122a-122n** assoziiert, der einen oder mehrere Parameter beinhaltet, von denen jeder das Verhalten des entsprechenden Subgraphen definiert.

[0030] Die Konstruktion eines Datenflussgraphen kann in manchen Fällen von sehr technischer Natur sein. Obwohl sie zum Erzielen spezieller Geschäftszwecke geschrieben ist, wird die zugrundeliegende

Struktur und Konstruktion des Graphen basierend auf technischen Überlegungen bestimmt. Zum Beispiel können Graphenkomponenten ausgewählt werden, um eine Wiederverwendbarkeit zu maximieren oder um eine Parallelverarbeitung zu unterstützen. Andererseits kann es hauptsächlich eine Geschäftsentscheidung sein, wie und wo ein Graph zu verwenden ist. Manche der Parameter, die mit einem parametrisierten Datenflussgraphen assoziiert sind, können verwendet werden, um zu ermöglichen, dass Geschäftsbenutzer Datenflussgraphen anpassen, ohne dass der Benutzer die technischen Komplexitäten hinter ihren Implementierungen verstehen muss. Die parametrisierten Datenflussgraphen vereinfachen eine Anpassung und fördern eine Widerverwendung.

[0031] Eine Oberfläche zur Identifikation von Parameterwerten zum Konstruieren eines Datenflussgraphen kann auf einer Client-Maschine präsentiert werden. Bei manchen Implementierungen kann der Client auf eine Entwicklungsumgebung, die auf einem Server läuft, unter Verwendung eines Web-Browsers auf dem Client, der die Parameteroberfläche bereitstellt, und unter Verwendung einer Skriptsprache, die eine gewisse Fähigkeit für eine clientseitige Verarbeitung bereitstellt, zugreifen. Die Skriptsprache kann mit dem Server kommunizieren, um Parameter zu aktualisieren und andere notwendige Operationen durchzuführen. Diese Kommunikation kann über eine Brückenmaschine stattfinden, die die Kommunikationen zwischen dem Client und dem Server übersetzt, der eine Entwicklungsumgebung ausführt, die Objekte und assoziierte Parameterwerte für die Graphen speichert, die konstruiert werden. Die Oberfläche ermöglicht, dass ein Benutzer die Parameter eines parametrisierten Datenflussgraphen konfiguriert, selbst wenn dem Benutzer das technische Wissen bezüglich Datenflussgraphen und einer Datenflussgraphenkonfiguration fehlt.

[0032] Eine Konfigurationsoberfläche, die auf einer Client-Vorrichtung präsentiert wird, ermöglicht, dass ein Benutzer auf das Graphenkonfigurationsmodul **16** aus **Fig. 1** zugreift. Durch die Konfigurationsoberfläche kann der Benutzer Charakteristiken der Datenquelle, der Datensenken und der durch den Datenflussgraphen durchzuführenden Transformationen spezifizieren, ohne technisches Wissen über eine Datenflussgraphenkonfiguration zu benötigen. Basierend auf den durch den Benutzer spezifizierten Charakteristiken können Parameterwerte für die Parametersätze **108**, **110**, **112a-112b**, **122a-122n** zugewiesen werden, wodurch das Verhalten des Datenflussgraphen gemäß den durch den Benutzer spezifizierten Charakteristiken definiert wird.

[0033] Innerhalb der Konfigurationsoberfläche können die Parameter der Parametersätze **108**, **110**, **112a-112n** kombiniert und neu in Gruppen zur Interaktion mit einem Benutzer organisiert werden, die z. B. Geschäftsüberlegungen anstelle technischer Überlegungen wiedergeben. Die Konfigurationsoberfläche zum Empfangen von Werten für die Parameter basierend auf einer Benutzereingabe kann unterschiedliche Parameter gemäß Beziehungen zwischen Parametern auf eine flexible Weise anzeigen, die nicht notwendigerweise durch Aspekte der Entwicklungsumgebung auf dem Server eingeschränkt ist. Ein Beispiel für eine Konfigurationsoberfläche ist in der US-Veröffentlich mit der Nr. 2011/0145748 beschrieben, deren Inhalte hiermit durch Bezugnahme in ihrer Gesamtheit aufgenommen sind.

[0034] Unter Bezugnahme auf **Fig. 3** ermöglicht eine Konfigurationsoberfläche **300**, dass ein Benutzer Parametern des Datenflussgraphen **100** Werte zuweist. Bei dem Beispiel aus **Fig. 3** sind die Parameter in eine erste Gruppe von Parametern unterteilt, die durch ein Quellensymbol **302**, das Parameter für die Datenquelle **102** repräsentiert, ein Transformationssymbol **304**, das Parameter für die Komponente **104** repräsentiert, und ein Senkensymbol **306a-306n**, die jeweils Parameter für eine entsprechende der Datensenken **106a-106n** repräsentierten, repräsentiert wird. Der Benutzer kann mit jedem der Symbole assoziierte Charakteristiken eingeben. Basierend auf den durch den Benutzer eingegebenen Charakteristiken können Werte für die Parameter für die Datenquelle **102**, die Datensenken **106a-106n** und die Komponente **104** bestimmt werden.

[0035] Bei dem Beispiel aus **Fig. 3** wird eine erste Charakteristik, die mit dem Quellensymbol **302** assoziiert ist, als ein Textfeld **308** angezeigt, in das der Benutzer einen Text eingeben kann. Der Typ von Text kann eingeschränkt sein (z. B. Integer, Gleitkommawerte, Strings oder andere Typen) oder kann einen beliebigen Typ von Text akzeptieren. Eine zweite Charakteristik wird als eine Drop-Down-Liste **310** mit vorbesetzten Werten angezeigt. Eine dritte Charakteristik wird als eine Optionsschaltfläche **312** angezeigt. Ein Feld **314** für eine vierte Charakteristik erlaubt einem Benutzer, eine größere Textmenge einzugeben. Andere Typen von Anzeigen können ebenfalls verwendet werden. Beispielsweise kann ein Benutzer dazu in der Lage sein, einen Dateiselektor zu öffnen, um einen speziellen Datensatz zur Verwendung als eine Datenquelle oder einen Ort, von dem aus auf einen Datensatz zuzugreifen oder dieser über ein Netz herunterzuladen ist, auszuwählen. Mit den Senkensymbolen **306a-306n** und dem Transformationssymbol **304** assoziierte Charakteristiken können ebenfalls eingegeben werden.

[0036] Die mit dem Quellensymbol **302** oder den Senkensymbolen **306a-306n** assoziierten Charakteristiken können eine Charakteristik beinhalten, die den Typ von Datensatz für die Datenquelle **102** oder eine der Datensenken **106a-106** angibt. Beispielsweise kann der Benutzer aufgefordert wer-

den, einen Typ eines Datensatzes aus einer Liste unterstützter Typen, wie etwa Dateien, Tabellen, HDFS(Hadoop Distributed File System -verteiltes Hadoop-Dateisystem)-Datensätze, HIVE-Tabellen, Warteschlangen oder andere Typen, auszuwählen. Mit Dateien sind Daten mit einer Form gemeint, die als eine Datei beschrieben oder mit einem Aufzeichnungsformat definiert werden kann, das mit der Implementierung des Datenflussgraphen **100** kompatibel ist. Beispielsweise sind MVS(Multiple Virtual Storage - mehrfache virtuelle Speicherung)-Dateien Beispiele für Dateien. Der Benutzer kann aufgefordert werden, einen Pfad oder einen Ort des Datensatzes zu spezifizieren. Andere Charakteristiken, für die der Benutzer aufgefordert wird, können von dem Typ von Datensatz abhängen, der von dem Benutzer ausgewählt wird. Beispielsweise kann der Benutzer Charakteristiken bezüglich eines Aufzeichnungsformats des Datensatzes, eines Schlüssels des Datensatzes, einer Konfigurationsdatei für einen Tabellendatensatz, eines Typs einer Kompression oder Verschlüsselung oder andere Charakteristiken eingeben. Wenn mehrere Datensenken verwendet werden, kann jede Datensenke ihren eigenen Typ von Datensatz und andere assoziierte Charakteristiken aufweisen.

[0037] Die mit dem Quellensymbol **302** assoziierten eingegebenen Charakteristiken werden verwendet, um die Datenquelle **102** automatisch zu parametrisieren, so dass der Datenflussgraph bei Ausführung Daten von der angemessenen Quelle empfängt. Gleichermaßen werden die mit jedem Quellensymbol **306a-306n** assoziierten eingegebenen Charakteristiken verwendet, um jede Datenquelle **106a-106n** automatisch zu parametrisieren, so dass eine Datenausgabe von dem Datenflussgraphen **100** bei Ausführung an das angemessene Ziel gesendet wird.

[0038] Die mit dem Transformationssymbol **304** assoziierten Charakteristiken können eine Charakteristik beinhalten, die den Typ einer Verarbeitung angeben, die an Daten von der Datenquelle durchzuführen ist. Beispielsweise kann der Benutzer aufgefordert werden, einen Typ eines Prozesses aus einer Bibliothek unterstützter Prozesse auszuwählen. Jeder Prozess entspricht einem zuvor konstruierten, gespeicherten Subgraphen **120a-120n**, der durch die Komponente **104** referenziert werden kann. Beispielsweise kann das System Prozesse unterstützen, wie etwa einen Zuordnungsprozess, einen Filterprozess, einen Deduplizierungsprozess (der manchmal als ein Dedupe-Prozess bezeichnet wird), einen Aggregationsprozess, einen Replikationsprozess, einen Normalisierungsprozess, einen Unterteilungs- und Sortierungsprozess, einen Datenerfassungsänderungsprozess, einen Zusammenführungsprozess oder andere Prozesse. Andere Charakteristiken, für die der Benutzer aufgefordert wird, können von dem Typ von Prozess abhängen. Beispielsweise kann der Benutzer für einen Zuordnungsprozess Zuordnungsregeln eingeben und einen Zieldatensatz für die Zuordnung identifizieren. Für einen Filterprozess kann der Benutzer Filterregeln eingeben. Für einen Dedupe-Prozess kann der Benutzer einen Dedupe-Schlüssel eingeben und Aufzeichnungsauswahlregeln spezifizieren. Für einen Zusammenführungsprozess kann der Benutzer einen Zusammenführungsdatensatz identifizieren und Zusammenführungsregeln definieren. Bei manchen Beispielen kann der Benutzer mehrere Prozesse auswählen und die gewünschte Sequenz der ausgewählten Prozesse angeben.

[0039] Die eingegebenen Charakteristiken, die mit den Transformationssymbol **304** assoziiert sind, werden verwendet, um die Komponente **104** sowie beliebige relevante Subgraphen **120a-120n** automatisch zu parametrisieren. Die Komponente **104** wird durch Referenzieren des (der) Subgraphen parametrisiert, der (die) den Typen von Prozessen entspricht (entsprechen), die von dem Benutzer ausgewählt sind. Jeder der referenzierten Subgraphen wird auch basierend auf den Charakteristiken parametrisiert, die von dem Benutzer für den entsprechenden Prozess eingegeben werden. Bei Ausführung wird der Datenflussgraph **100** jeden der Subgraphen **120a-120n** ausführen, die durch die Komponente referenziert werden, wobei das Verhalten jedes referenzierten Subgraphen durch die Parametrisierung basierend auf den vom Benutzer eingegebenen Charakteristiken definiert wird. Beispielswiese kann der Datenflussgraph **100** jeden der referenzierten Subgraphen **120a-120n** in einem seriellen Pipeline-Prozess ausführen, in dem keine Daten während einer Ausführung der Subgraphen gespeichert werden.

[0040] Bei den hier beschriebenen Beispielen unterstützt das System eine einzige Datenquelle und mehrere Datensenken, wie etwa bis zu zwei Datensenken, fünf Datensenken, zehn Datensenken, zwanzig Datensenken oder eine andere Anzahl an Datensenken. Die Parameter für jede Datensenke können getrennt definiert werden, z. B. so, dass jede Datensenke ein anderer Typ von Datensatz sein kann und/oder ihren eigenen Satz von Parametern aufweisen kann. Bei manchen Beispielen kann das System mehrere Datenquellen unterstützen, z. B. können Daten von mehreren Quellen unter Verwendung eines Zusammenführungsprozesses, eines Nachschlageprozesses oder anderer Typen von Prozessen verarbeitet werden. Beispielsweise können mehrere Quellen durch einen Zusammenführungsprozess, einen Nachschlageprozess oder eine Kombination von sowohl Zusammenführungs- als auch Nachschlageprozessen in eine effektive einzige Quelle zusammengeführt werden. Bei manchen Beispielen kann das System eine einzige Datensenke unterstützen.

[0041] Bei manchen Beispielen kann das System automatisch eine Konfiguration basierend auf einer

Benutzereingabe erzeugen, die Einzelheiten für eine Datenquelle und einen Satz von Verarbeitungsschritten bereitstellt. Beispielsweise kann der Benutzer Einzelheiten einer Datenquelle und eines Satzes von Verarbeitungsschritten durch eine Benutzeroberfläche, wie etwa ein Spreadsheet, bereitstellen und das System kann ein Computerprogramm automatisch basierend auf den durch den Benutzer eingegebenen Einzelheiten konfigurieren.

[0042] Unter Bezugnahme auf **Fig. 4** werden in einem Beispielprozess zum Konfigurieren eines ersten computerausführbaren Programms Informationen, die eine Quelle von Daten und ein Datenziel angeben, empfangen (**400**). Beispielsweise können die Informationen den Typ eines Datensatzes, einen Ort oder Pfad für den Datensatz und andere Charakteristiken über den Datensatz angeben. Eine Charakterisierung eines Prozesses wird empfangen (**402**). Die Charakterisierung beinhaltet einen Typ des Prozesses und Werte für jeweilige Charakteristiken des Prozesses.

[0043] Werte werden basierend auf den empfangenen Informationen automatisch zu jeweiligen Parametern des ersten computerausführbaren Programms zugewiesen (**404**), um zu bewirken, dass das erste computerausführbare Programm Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt, wenn das Programm ausgeführt wird. Beispielsweise kann das erste computerausführbare Programm ein Datenflussgraph sein.

[0044] Das erste computerausführbare Programm wird basierend auf der empfangenen Charakterisierung des Prozesses automatisch dazu konfiguriert, ein zweites computerausführbares Programm zu referenzieren (**406**). Das Konfigurieren beinhaltet Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses (**408**). Das Konfigurieren beinhaltet auch Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind (**410**). Beispielsweise kann das zweite computerausführbare Programm ein Subgraph sein.

[0045] **Fig. 5A-5G** sind Beispiele für eine Konfigurationsoberfläche, durch die ein Benutzer Informationen über eine Datenquelle oder ein Datenziel eingeben kann. Unter Bezugnahme auf **Fig. 5A** wird der Benutzer aufgefordert, den Typ des Quellen- oder Zieldatensatzes auszuwählen.

[0046] Unter Bezugnahme auf **Fig. 5B** und **Fig. 5C** wird der Benutzer, falls der Typ des Datensatzes eine Datei ist, aufgefordert, Validierungsdetails und Charakteristiken einzugeben, wie etwa die URL der Datei, das Aufzeichnungsformat der Datei, ob die Datei komprimiert oder verschlüsselt oder beides ist, ob die Datei gelesen oder geschrieben werden soll, einen natürlichen Schlüssel der Datei, einen Datenreihenfolgenschlüssel der Datei und/oder andere Parameter.

[0047] Unter Bezugnahme auf **Fig. 5D** können Metadaten über einen Datensatz eingegeben werden, wie etwa ein RPfad (rpath) oder Repositoriumname einer physischen Datei.

[0048] Unter Bezugnahme auf **Fig. 5E** wird der Benutzer, falls der Typ des Datensatzes eine Tabelle ist, aufgefordert, Charakteristiken einzugeben, wie etwa, ob eine Tabelle oder eine Abfrage (z. B. eine SQL-Abfrage) verwendet werden soll, ob die Tabelle geschrieben oder gelesen werden soll, eine Datenbankkonfigurationsdatei für die Tabelle, einen natürlichen Schlüssel der Tabelle, einen Datenreihenfolgenschlüssel der Tabelle oder andere Parameter. Falls die Abfrage ausgewählt wird, kann der Benutzer die Abfrage spezifizieren.

[0049] Unter Bezugnahme auf **Fig. 5F** wird der Benutzer, falls der Datensatz ein HDFS-Datensatz ist, aufgefordert, Charakteristiken einzugeben, die den HDFS-Datensatz beschreiben, wie etwa, ob die HDFS-Datei geschrieben oder gelesen werden soll, den Pfad des HDFS-Dateisystems, den Namen der HDFS-Datei, ein Suffix der HDFS-Datei (falls vorhanden), einen Pfad zu einer Aufzeichnungsformatdatei und/oder beliebige einer Vielzahl anderer Charakteristiken. Der Benutzer kann Charakteristiken bereitstellen, die den Dateityp oder das Dateiformat angeben, wie etwa den Typ einer zu lesenden oder zu schreibenden HDFS-Datei, wie etwa einer nichtinterpretierten Datei (z. B. unkomprimierter nichtformatierter Text (z. B. ASCII oder UTF8) oder Binärdaten), ein spaltenorientiertes Zeilenformat, wie etwa das Apache-Optimized-Row-Columnar-Format, ein spaltenorientiertes Speicherungsformat, wie etwa Apache Parquet, eine Binärdatei einschließlich Schlüssel-Wert-Paaren, wie etwa einer Hadoop-Sequenz-Datei, ein Binärdateiformat, wie etwa Apache Avro, Textdaten oder andere Typen. Der Benutzer kann Charakteristiken bereitstellen, die den Typ einer Kompression angeben, wie etwa gzip, Snappy, Deflate, bzip2, zlib, LZ4, LZO oder andere Typen einer Kompression. Der Benutzer kann das Aufzeichnungsformat für den HDFS-Datensatz bereitstellen, das direkt aus einem existierenden HDFS-Datensatz abgeleitet werden kann, oder kann das Aufzeichnungsformat bereitstellen. Bei manchen Beispielen kann der Benutzer nach zusätzlichen Charakteristiken in Abhängigkeit von dem Typ einer Datei, dem Typ einer Kompression oder beidem aufgefordert werden.

[0050] Unter Bezugnahme auf **Fig. 5G** wird der Benutzer, falls der Datensatz eine HIVE-Tabelle ist, auf-

gefordert, Charakteristiken einzugeben, die die HIVE-Tabelle beschreiben, wie etwa, ob die HIVE-Tabelle geschrieben oder gelesen werden soll, das Dateiformat der HIVE-Tabelle, das Aufzeichnungsformat, den Pfad einer Aufzeichnungsformatdatei, die Datenbank, ob die Tabelle existiert, die Tabelle und/oder beliebige einer Vielzahl anderer Charakteristiken. Bei manchen Beispielen kann der Benutzer nach zusätzlichen Charakteristiken in Abhängigkeit von der Auswahl des Typs einer Tabelle und/oder des Typs einer Kompression aufgefordert werden.

[0051] **Fig. 6** ist ein Beispiel für eine Konfigurationsoberfläche **600**, in der ein Benutzer Verarbeitungsschritte definieren kann. Bei dem Beispiel aus **Fig. 6** wurden ein Quellendatensatz **602** und zwei Zieldatensätze **604a**, **604b** bereits definiert. Ein Verarbeitungsschrittfenster **606** ermöglicht, dass der Benutzer die Verarbeitungsschritte auswählt, die an dem Quellendatensatz durchzuführen sind.

[0052] **Fig. 7** ist ein Beispiel für eine Konfigurationsoberfläche **700**, die einen Feed zeigt, in dem eine Reihe von Verarbeitungsschritten durch den Benutzer spezifiziert wurde. Ein Quellendatensatz **702** wurde definiert. Vier Verarbeitungsschritte wurden ebenfalls definiert: ein Zuordnungsschritt **704**, ein Dedupe-Schritt **706**, ein zweiter Zuordnungsschritt **708** und ein Zusammenführungsschritt **710**. Verschiedene Charakteristiken für jeden Prozess wurden spezifiziert.

[0053] Wenn ein Datenflussgraph basierend auf dem in der Konfigurationsoberfläche **700** gezeigten Feed parametrisiert wird, wird die Datenquelle dazu konfiguriert, Daten von der Quellentabelle Teradata_cc_trans zu empfangen, wie in dem Quellendatensatzfenster **702** spezifiziert ist. Die Komponente des Datenflussgraphen wird dazu konfiguriert, den Zuordnungssubgraphen, den Dedupe-Subgraphen, wieder den Zuordnungssubgraphen und den Zusammenführungssubgraphen zu referenzieren. Jeder referenzierte Subgraph wird gemäß den in dem Fenster **704-710** gezeigten Charakteristiken für den jeweiligen Prozess parametrisiert.

[0054] Wie aus dem Feed in der Konfigurationsoberfläche **700** gesehen werden kann, kann ein Benutzer Charakteristiken zur Konfiguration eines Datenflussgraphen eingeben, ohne ein detailliertes technisches Verständis einer Graphenkonfiguration haben zu müssen. Beispielsweise kennt der Benutzer bei dem Beispiel aus **Fig. 7** die Quelle der Daten, die er verarbeiten will, und das Ziel (nicht gezeigt) für die verarbeiteten Daten. Der Benutzer weiß, dass er die Quellendaten zu sich selbst zuordnen will, z. B. um die Zuordnung eines oder mehrerer Felder zu ändern. Er weiß, dass er die zugeordneten Daten unter Verwendung des Dedupe-Schlüssels {Menge} deduplizieren (dedupe) und die erste Aufzeichnung in jeder Gruppe behalten will. Nach dem Dedupe-Schritt will er die Daten wieder sich selbst zuordnen. Schließlich will er die deduplizierten Daten und die zugeordneten Daten zu der Tabelle mixed_cc_accounts zusammenführen. Das heißt, um diesen Prozessfluss einzurichten, kann der Benutzer das allgemeine Datenverarbeitungsgerüst verstehen, aber muss nicht wissen, wie ein Datenflussgraph zu konfigurieren ist.

[0055] Bei manchen Beispielen kann der Benutzer durch die Konfigurationsoberfläche **700** einen Satz von Testdaten auswählen, der als der Quellendatensatz **702** zu verwenden ist. Anstatt den Datenflussgraphen unter Verwendung eines extrem großen Datensatzes (z. B. eines Datensatzes mit tatsächlichen Kundenaufzeichnungen) zu testen, kann der Benutzer beispielweise einen kleineren Satz von Daten auswählen, der schneller verarbeitet werden kann. Bei manchen Beispielen kann der Benutzer, z. B. zu Testzwecken, ein Ziel für die verarbeiteten Daten auswählen, das sich von dem Ziel unterscheidet, zu dem der Datenflussgraph seine verarbeiteten Daten senden wird, wenn er sich tatsächlich in Betrieb befindet. Beispielsweise kann der Benutzer ein Testdatenziel auswählen, so dass ein Datensatz mit tatsächlichen Kundenaufzeichnungen nicht überschrieben wird. Unabhängig von dem Quellendatensatz **702** und dem Ziel der verarbeiteten Daten verbleibt die Konfiguration der anderen Prozesse **706-708** unverändert und dementsprechend ist es für den Benutzer unkompliziert, den Quellendatensatz und/oder das Ziel wie gewünscht zu aktualisieren.

[0056] Bei manchen Beispielen kann die Konfigurationsoberfläche **700** Angaben darüber bereitstellen, ob die durch den Benutzer eingegebenen Parameter Sinn ergeben. Beispielsweise können durch den Benutzer eingegebene Verarbeitungsschlüssel gegeneinander validiert werden. Bei einem Beispiel kann, falls ein Benutzer einen Datenflussgraphen erzeugt, der einen Datensatz unter Verwendung eines ersten Schlüssels dedupliziert und die resultierenden deduplizierten Daten unter Verwendung eines zweiten Schlüssels sortiert, ein Alarm in der Konfigurationsoberfläche angezeigt werden.

[0057] Bei manchen Beispielen kann das System, nachdem der Benutzer den Datenflussgraphen konfiguriert hat, bestimmen, wie der Datenflussgraph auszuführen ist, z. B. um eine Verarbeitungszeit oder Rechenkosten zu reduzieren. Beispielsweise kann das System für einen gegebenen Prozess bestimmen, ob die Daten basierend auf Parametern eines oder mehrerer vorgelagerter Prozesse, wie etwa in vorgelagerten Prozessen verwendeten Schlüsseln, neu zu unterteilen sind. Bei einem Beispiel kann das System, falls ein ähnlicher Schlüssel in einem vorgelagerten Prozess verwendet wird, bestimmen, dass eine Neuunterteilung der Daten mit dem gleichen Schlüssel in einem nachgelagerten Prozess überflüssig ist. Das

Bestimmen, ob eine Neuunterteilung oder Sortierung angemessen sind, kann automatisch ohne Benutzereingabe durchgeführt werden.

[0058] Bei manchen Beispielen können Informationen, die die Leistungsfähigkeit des Datenflussgraphen angeben, während einer Ausführung des Datenflussgraphen in der Konfigurationsoberfläche **700** angezeigt werden. Die Informationen können Informationen beinhalten, die die Anzahl an Aufzeichnungen in jeden Verarbeitungsschritt und aus jedem Verarbeitungsschritt angeben. Beispielsweise kann der Dedupe-Schritt **706** die Anzahl an eingehenden Aufzeichnungen, die Anzahl an ausgehenden Aufzeichnungen und die Anzahl an Aufzeichnungen, die als Duplikate verworfen werden, aufzeichnen und anzeigen. Der Zusammenführungsschritt **710** kann die Anzahl an eingehenden Aufzeichnungen von einem ersten Datensatz, die Anzahl an eingehenden Aufzeichnungen von einem zweiten Datensatz, die Anzahl an ausgehenden Aufzeichnungen und die Anzahl an verworfenen Aufzeichnungen von jedem Datensatz aufzeichnen und anzeigen. Diese Informationen können dem Benutzer eine schnelle Rückmeldung bereitstellen, die z. B. angibt, ob er den Datenflussgraphen auf die Weise konfiguriert hat, die er beabsichtigt hat, wodurch ein Weg für einen nichttechnischen Benutzer zur Fehlerbehebung und zum Debugging des Datenflussgraphen bereitgestellt wird. Beispielsweise kann unter Bezugnahme auf **Fig. 8** ein nichttechnischer Benutzer, der sieht, dass 115.911 Aufzeichnungen in einen Dedupe-Schritt gelesen wurden, von denen 108.191 als Duplikate verworfen wurden, diese Information verwenden, um zu bestimmen, dass ein Parameter, wie etwa der Dedupe-Schlüssel, möglicherweise inkorrekt konfiguriert wurde.

[0059] Bei manchen Beispielen kann die Konfigurationsoberfläche **700** als eine Spreadsheet-Oberfläche präsentiert werden, durch die der Benutzer einen Datenflussgraphen aufbauen und konfigurieren kann.

[0060] Bei manchen Beispielen können Quellendatensätze, Zieldatensätze oder beides erzeugt werden, z. B. basierend auf Charakteristiken, die zu verwendende Daten identifizieren, wie etwa Daten in Spreadsheets, Datenbanken oder anderen Orten.

[0061] Unter Bezugnahme auf **Fig. 9A** und **Fig. 9B** kann bei manchen Beispielen ein zielspezifischer Prozess, wie etwa ein Zuordnungs- bzw. Filterprozess, angewandt werden, bevor die verarbeiteten Daten von dem Datenflussgraphen zu dem Datenziel gesendet werden.

[0062] Unter Bezugnahme auf **Fig. 10** kann ein Client-System **802** dem Benutzer eine Konfigurationsoberfläche **804** anzeigen. Ein Parametersatz **814**, der basierend auf Interaktionen mit dem Benutzer durch die Konfigurationsoberfläche **804** erzeugt wird, kann auf einem Server **808** gespeichert werden. Folglich werden Änderungen, die durch die Konfigurationsoberfläche **804** vorgenommen werden, von dem Client **802** über eine Brücke **806** an den Server **808** gesendet. Der Client **802** sendet, wie durch den Pfeil **820** repräsentiert, eine Nachricht an die Brücke **806** in einem Format, zum Beispiel eine Nachricht, die unter Verwendung eines Simple-Object-Access-Protocol (SOAP - Einfacher-Objektzugriff-Protokoll) gesendet wird. Die Brücke **806** übersetzt die Nachricht in ein neues Format und beginnt, falls notwendig, eine Client-Sitzung mit dem Server **808**. Die Brücke **806** sendet, wie durch den Pfeil **822** repräsentiert, eine Nachricht an den Server **808** in einem Format, das von dem Server **808** verstanden wird, zum Beispiel eine COM+-Nachricht. Der Server **808** empfängt die Nachricht und aktualisiert den Parametersatz. Der Server **808** sendet, wie durch den Pfeil **824** repräsentiert, eine Antwort an die Brücke **806**, die beliebige Änderungen enthält, die während der von dem Client **802** empfangenen Eingabe an dem Parameter aufgetreten sind. Die Brücke **806** decodiert die Nachricht und erzeugt eine Antwortnachricht für den Client **802**. Die Brücke **806** sendet, wie durch den Pfeil **826** repräsentiert, die Antwortnachricht an den Client **802**. Der Client **802** aktualisiert die Konfigurationsoberfläche **804**, um die Änderungen wiederzugeben, einschließlich Anzeigen beliebiger Komponenten, die zuvor aufgrund des Fehlers einer Vorbedingung, wie oben beschrieben, verdeckt waren.

[0063] Der Benutzer kann dem Client **802** auch angeben, dass er den Datenflussgraphen, der konstruiert wird, unter Verwendung von Probendaten basierend auf dem aktuellen Satz von Parametern, der vollständig sein kann oder nicht, ausführen möchte. Wie oben, sendet der Client **802** eine Nachricht über die Brücke **806** an den Server **808**. Der Server **808** wendet beliebige Änderungen auf den Parametersatz an und ein Prozess **816**, der auf dem Server läuft, kompiliert den Datenflussgraphen. Der kompilierte Datenflussgraph akzeptiert Daten von den Probendatensätzen **810**, **812** und führt den kompilierten Datenflussgraphen aus. Der Datenflussgraph produziert die angeforderte Ausgabe in einen Ausgabedatensatz **818**. Die Ausgabe des Datenflussgraphen sind die Zwischendaten, die von dem Client **802** angefordert werden, und nicht notwendigerweise die Daten, die durch eine vollständige Ausführung des Datenflussgraphen produziert würden.

[0064] Wie oben beschrieben, werden die resultierenden Daten von dem Server **808** über die Brücke **806** an den Client **802** gesendet.

[0065] Der oben beschriebe Graphenkonfigurationsansatz kann unter Verwendung von Software zur Ausführung auf einem Computer implementiert werden. Beispielsweise bildet die Software zum Bei-

spiel Prozeduren in einem oder mehreren Computerprogrammen, die auf einem oder mehreren programmierten oder programmierbaren Computersystemen ausgeführt werden (die verschiedene Architekturen aufweisen können, wie etwa verteilt, Client/Server oder Gitter), die jeweils wenigstens einen Prozessor, wenigstens ein Datenspeicherungssystem (einschließlich eines flüchtigen und nichtflüchtigen Speichers und/oder Speicherungselementen), wenigstens eine(n) Eingabevorrichtung oder -port und wenigstens eine Ausgabevorrichtung oderport beinhalten. Die Software kann ein oder mehrere Module eines größeren Programms bilden, das zum Beispiel andere Dienste in Zusammenhang mit der Gestaltung und Konfiguration von Berechnungsgraphen bereitstellt. Die Knoten und Elemente des Graphen können als Datenstrukturen, die in einem computerlesbaren Medium gespeichert werden, oder andere organisierte Daten, die einem Datenmodell entsprechen, das in einem Datenrepositorium gespeichert ist, implementiert werden.

[0066] Die Software kann auf einem Speicherungsmedium, wie etwa einer CD-ROM, das durch einen Mehrzweckcomputer oder einen programmierbaren Spezialcomputer lesbar ist, bereitgestellt werden oder über ein Kommunikationsmedium eines Netzes an den Computer (in einem propagierten Signal codiert) geliefert werden, wo sie ausgeführt wird. Alle der Funktionen können auf einem Spezialcomputer oder unter Verwendung von Spezialhardware, wie etwa Koprozessoren, durchgeführt werden. Die Software kann auf eine verteilte Weise implementiert werden, bei der unterschiedliche Teile der Berechnung, die durch die Software spezifiziert werden, durch unterschiedliche Computer durchgeführt werden. Jedes solche Computerprogramm wird vorzugsweise auf Speicherungsmedien oder einer Speicherungsvorrichtung (z. B. Festkörperspeicher oder -medien oder magnetischen oder optischen Medien), die durch einen programmierbaren Allgemein- oder Spezialcomputer lesbar sind/ist, zum Konfigurieren und Betreiben des Computers gespeichert oder auf dieses heruntergeladen, wenn die Speicherungsmedien oder Speicherungsvorrichtung durch den Computer gelesen werden/wird, um die hier beschriebenen Prozeduren durchzuführen. Es kann auch erachtet werden, dass das erfindungsgemäße System als ein computerlesbares Speicherungsmedium implementiert ist, das mit einem Computerprogramm konfiguriert wird, wobei das so konfigurierte Speicherungsmedium bewirkt, dass ein Computersystem auf eine spezielle und vordefinierte Weise arbeitet, um die hier beschriebenen Funktionen durchzuführen.

[0067] Eine Anzahl an Ausführungsformen der Erfindung wurde beschrieben. Trotzdem versteht es sich, dass verschiedene Modifikationen vorgenommen werden können, ohne von dem Wesen und dem Schutzumfang der Erfindung abzuweichen. Zum Beispiel können manche der oben beschriebenen Schritte unabhängig von der Reihenfolge sein und können somit in einer Reihenfolge durchgeführt werden, die sich von der beschriebenen unterscheidet.

[0068] Andere Implementierungen liegen ebenfalls innerhalb des Schutzumfangs der folgenden Ansprüche.

ZITATE ENTHALTEN IN DER BESCHREIBUNG

Diese Liste der vom Anmelder aufgeführten Dokumente wurde automatisiert erzeugt und ist ausschließlich zur besseren Information des Lesers aufgenommen. Die Liste ist nicht Bestandteil der deutschen Patent- bzw. Gebrauchsmusteranmeldung. Das DPMA übernimmt keinerlei Haftung für etwaige Fehler oder Auslassungen.

Zitierte Patentliteratur

- US 62/516396 [0001]
- US 15/997278 [0001]
- US 2007/0011668 [0024]

Patentansprüche

1. Verfahren zum Konfigurieren eines ersten computerausführbaren Programms, wobei das Verfahren Folgendes beinhaltet:
Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche;
Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet;
automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und
automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren. wobei das Konfigurieren Folgendes beinhaltet:
Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und
Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

2. Verfahren nach Anspruch 1, wobei das erste computerausführbare Programm einen ersten Datenflussgraphen beinhaltet, der einen oder mehrere Knoten, die Datenverarbeitungskomponenten repräsentieren, und eine oder mehrere Verknüpfungen, die Flüsse von Daten zwischen den Datenverarbeitungskomponenten repräsentieren, beinhaltet.

3. Verfahren nach Anspruch 1 oder 2, wobei das zweite computerausführbare Programm einen zweiten Datenflussgraphen beinhaltet, der einen oder mehrere Knoten, die Datenverarbeitungskomponenten repräsentieren, und eine oder mehrere Verknüpfungen, die Flüsse von Daten zwischen den Datenverarbeitungskomponenten repräsentieren, beinhaltet.

4. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Informationen, die die Quelle von Daten angeben, einen Typ der Quelle von Daten identifizieren.

5. Verfahren nach Anspruch 4, wobei das Empfangen der Informationen, die die Quelle von Daten angeben, Empfangen einer Auswahl des Typs der Quelle von Daten aus einem Satz von einem oder mehreren Typen beinhaltet.

6. Verfahren nach Anspruch 5, wobei der Satz von einem oder mehreren Typen eine Datei, eine Tabelle, einen HDFS-Datensatz, eine HIVE-Tabelle und/oder eine Warteschlange beinhaltet.

7. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Informationen, die die Quelle von Daten angeben, Informationen beinhalten, die einen Ort der Quelle von Daten angeben.

8. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Informationen, die das Datenziel angeben, einen Typ des Datenziels identifizieren.

9. Verfahren nach Anspruch 8, wobei das Empfangen der Informationen, die das Datenziel angeben, Empfangen einer Auswahl des Typs des Datenziels aus einem Satz von einem oder mehreren Typen beinhaltet.

10. Verfahren nach Anspruch 9, wobei der Satz von einem oder mehreren Typen eine Datei, eine Tabelle, einen HDFS-Datensatz, eine HIVE-Tabelle und/oder eine Warteschlange beinhaltet.

11. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Informationen, die das Datenziel angeben, Informationen beinhalten, die mehrere Datenziele angeben.

12. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Informationen, die das Datenziel angeben, Informationen beinhalten, die einen Ort des Datenziels angeben.

13. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Charakterisierung des Prozesses eine Charakterisierung von jedem von mehreren Prozessen beinhaltet.

14. Verfahren nach einem der vorhergehenden Ansprüche, das Bestimmen der Charakteristiken, die mit dem Prozess assoziiert sind, basierend auf dem Typ des Prozesses beinhaltet.

15. Verfahren nach einem der vorhergehenden Ansprüche, wobei das Empfangen einer Charakterisierung eines Prozesses Empfangen einer Auswahl eines speziellen Prozesses aus einem Satz von einem oder mehreren Prozessen beinhaltet.

16. Verfahren nach Anspruch 15, wobei die Prozesse des Satzes einen Zuordnungsprozess, einen Filterprozess, einen Deduplizierungsprozess, einen Aggregationsprozess, einen Replikationsprozess, einen Normalisierungsprozess, einen Unterteilungs- und Sortierungsprozess, einen Datenerfas-

sungsänderungsprozess und/oder einen Zusammenführungsprozess beinhalten.

17. Verfahren nach einem der vorhergehenden Ansprüche, das automatisches Konfigurieren des ersten Datenflussgraphen dazu beinhaltet, den zweiten Datenflussgraphen zu einer Laufzeit des ersten Datenflussgraphen zu referenzieren.

18. Verfahren nach einem der vorhergehenden Ansprüche, wobei das automatische Konfigurieren des ersten Datenflussgraphen dazu, den zweiten Datenflussgraphen zu referenzieren, Konfigurieren des ersten Datenflussgraphen dazu beinhaltet, den zweiten Datenflussgraphen während einer Ausführung des ersten Datenflussgraphen auszuführen.

19. Nichtflüchtiges computerlesbares Speicherungsmedium, das Anweisungen zum Veranlassen eines Rechensystems dazu speichert, ein erstes computerausführbares Programm zu konfigurieren, wobei die Anweisungen Anweisungen beinhalten, um das Rechensystem zu Folgendem zu veranlassen:
Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche;
Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet;
automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und
automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren, wobei das Konfigurieren Folgendes beinhaltet:
Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und
Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

20. Rechensystem zum Konfigurieren eines ersten computerausführbaren Programms, wobei das Rechensystem Folgendes beinhaltet:
einen oder mehrere Prozessoren und einen Speicher, wobei der eine oder die mehreren Prozessoren und der Speicher zu Folgendem konfiguriert sind:
Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche;
Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet;
automatisches Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und
automatisches Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren, wobei das Konfigurieren Folgendes beinhaltet:
Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und
Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basierend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

21. Rechensystem zum Konfigurieren eines ersten computerausführbaren Programms, wobei das Rechensystem Folgendes beinhaltet:
ein Mittel zum Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben, durch eine Benutzeroberfläche;
ein Mittel zum Empfangen einer Charakterisierung eines Prozesses durch die Benutzeroberfläche, wobei die Charakterisierung einen Typ des Prozesses und Werte für jeweilige Charakteristiken, die mit dem Prozess assoziiert sind, beinhaltet;
ein Mittel zum automatischen Zuweisen von Werten zu jeweiligen Parametern des ersten computerausführbaren Programms basierend auf den empfangenen Informationen, um zu bewirken, dass das erste computerausführbare Programm, wenn es ausgeführt wird, Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt; und
ein Mittel zum automatischen Konfigurieren des ersten computerausführbaren Programms basierend auf der empfangenen Charakterisierung des Prozesses dazu, ein zweites computerausführbares Programm zu referenzieren, wobei das Konfigurieren Folgendes beinhaltet:
Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ des Prozesses; und
Zuweisen von Werten zu jeweiligen Parametern des zweiten computerausführbaren Programms basie-

rend auf den Werten für die jeweiligen Charakteristiken, die mit dem Prozess assoziiert sind.

Es folgen 14 Seiten Zeichnungen

Anhängende Zeichnungen

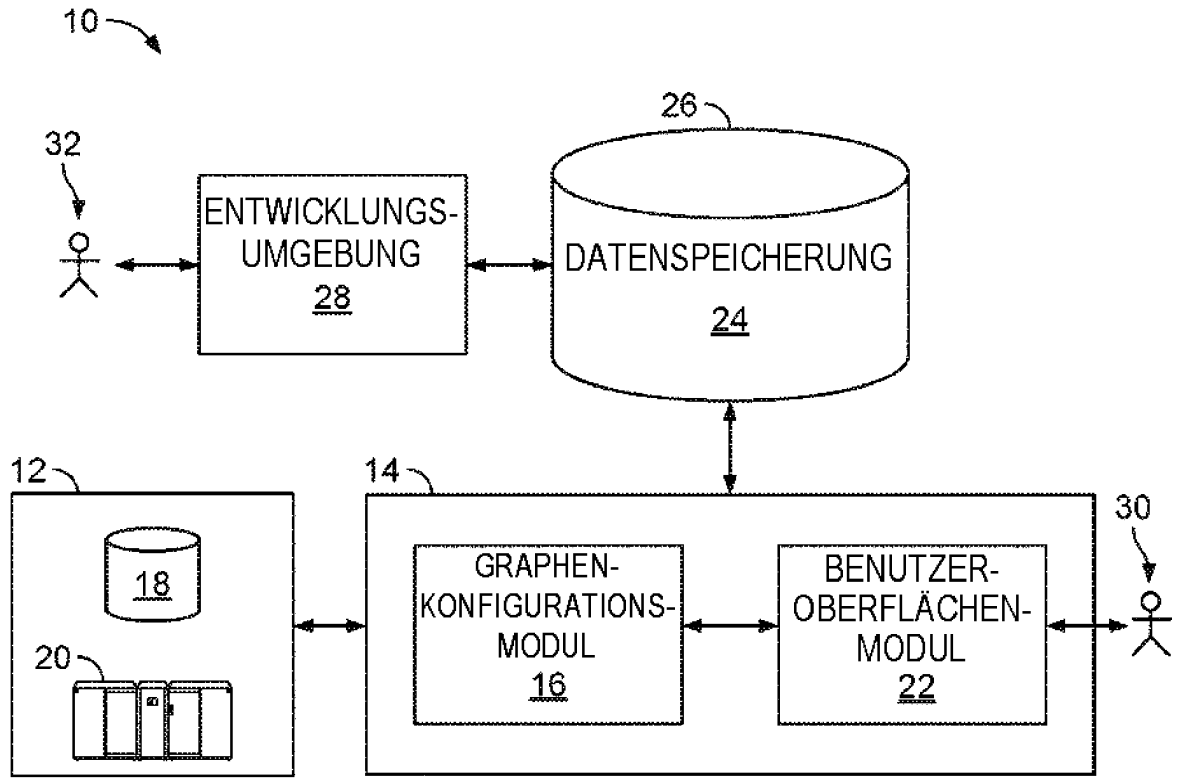


FIG. 1

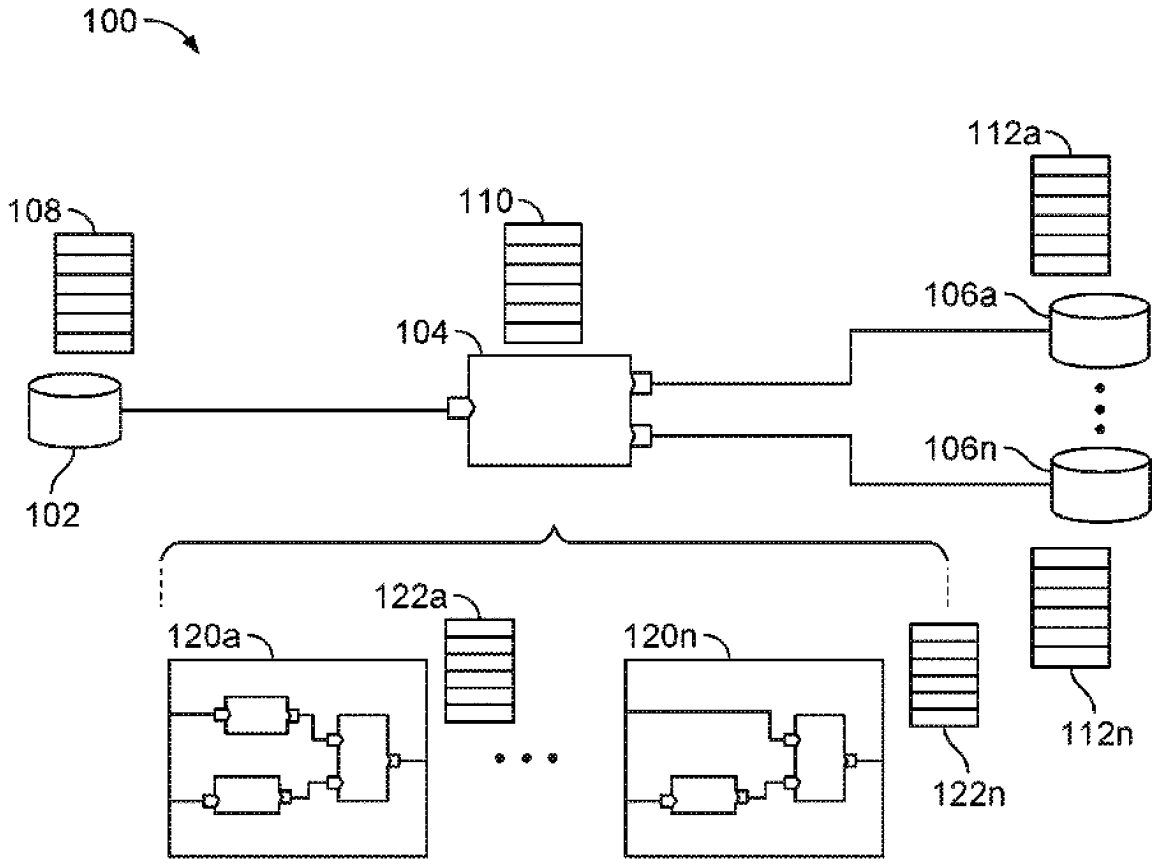


FIG. 2

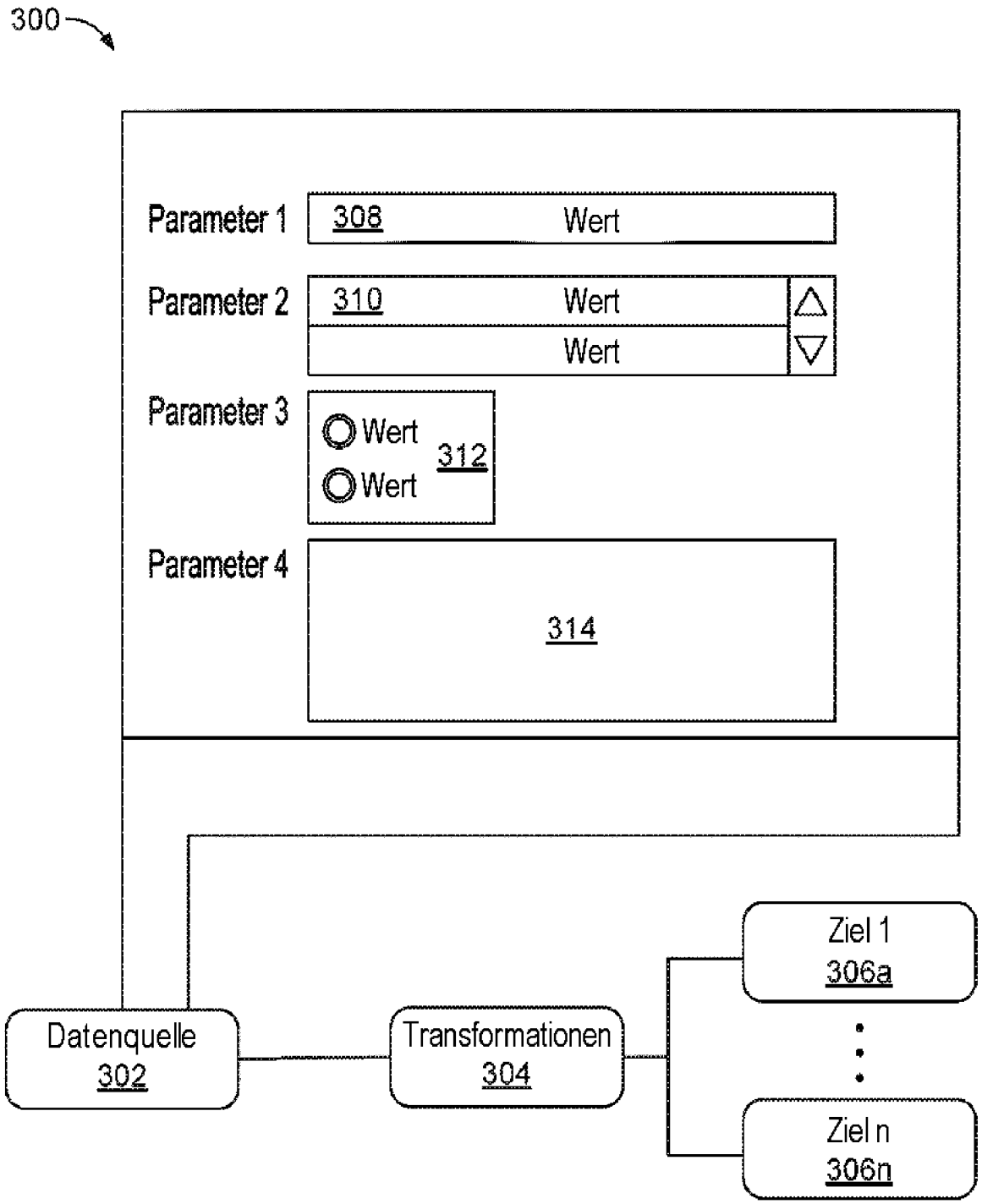

300
Parameter 1
308
Wert
Parameter 2
310
Wert
Wert
Parameter 3
Wert
Wert
312
Parameter 4
314
Datenquelle
302
Transformationen
304
Ziel 1
306a
Ziel n
306n

FIG. 3

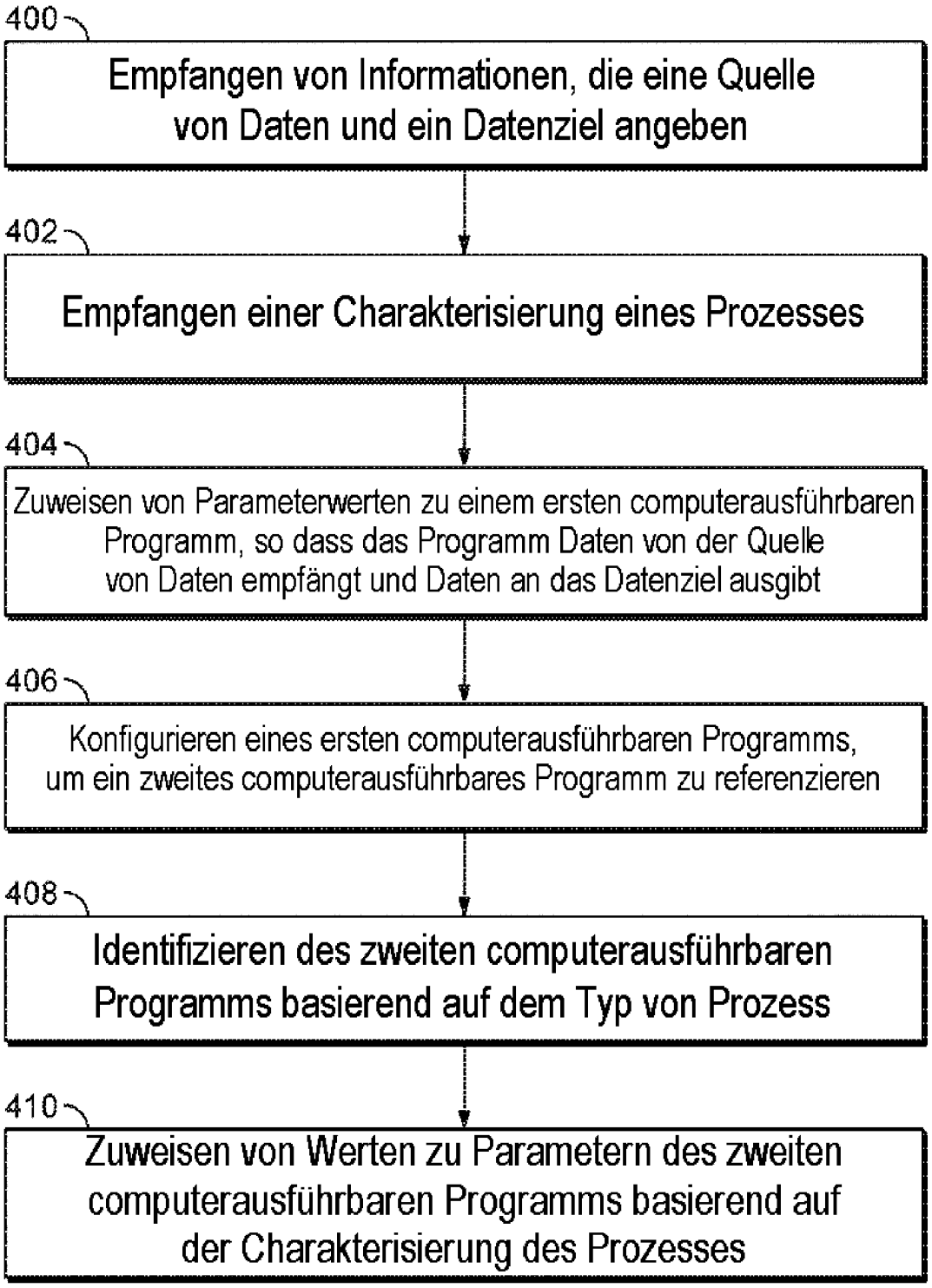

400
Empfangen von Informationen, die eine Quelle von Daten und ein Datenziel angeben
402
Empfangen einer Charakterisierung eines Prozesses
404
Zuweisen von Parameterwerten zu einem ersten computerausführbaren Programm, so dass das Programm Daten von der Quelle von Daten empfängt und Daten an das Datenziel ausgibt
406
Konfigurieren eines ersten computerausführbaren Programms, um ein zweites computerausführbares Programm zu referenzieren
408
Identifizieren des zweiten computerausführbaren Programms basierend auf dem Typ von Prozess
410
Zuweisen von Werten zu Parametern des zweiten computerausführbaren Programms basierend auf der Charakterisierung des Prozesses

FIG. 4


Datensatzeigenschaften
Datensatztyp auswählen:

FIG. 5A


Datensatzeigenschaften
Datensatztyp:
Datei
Konfigurieren der Datei
Physische Metadaten...
Organisationsmetadaten...
Validierungsprobleme
Fehler:
„Datei-URL“ ist leer. Klicken Sie auf „Physische Metadaten“ zum Durchsuchen, um eine Datei auszuwählen.
„Dateiaufzeichnungsformat“ ist leer. Klicken Sie auf „Physische Metadaten“ zum Durchsuchen, um eine Datei auszuwählen.
Daten ansehen und profilieren

FIG. 5B

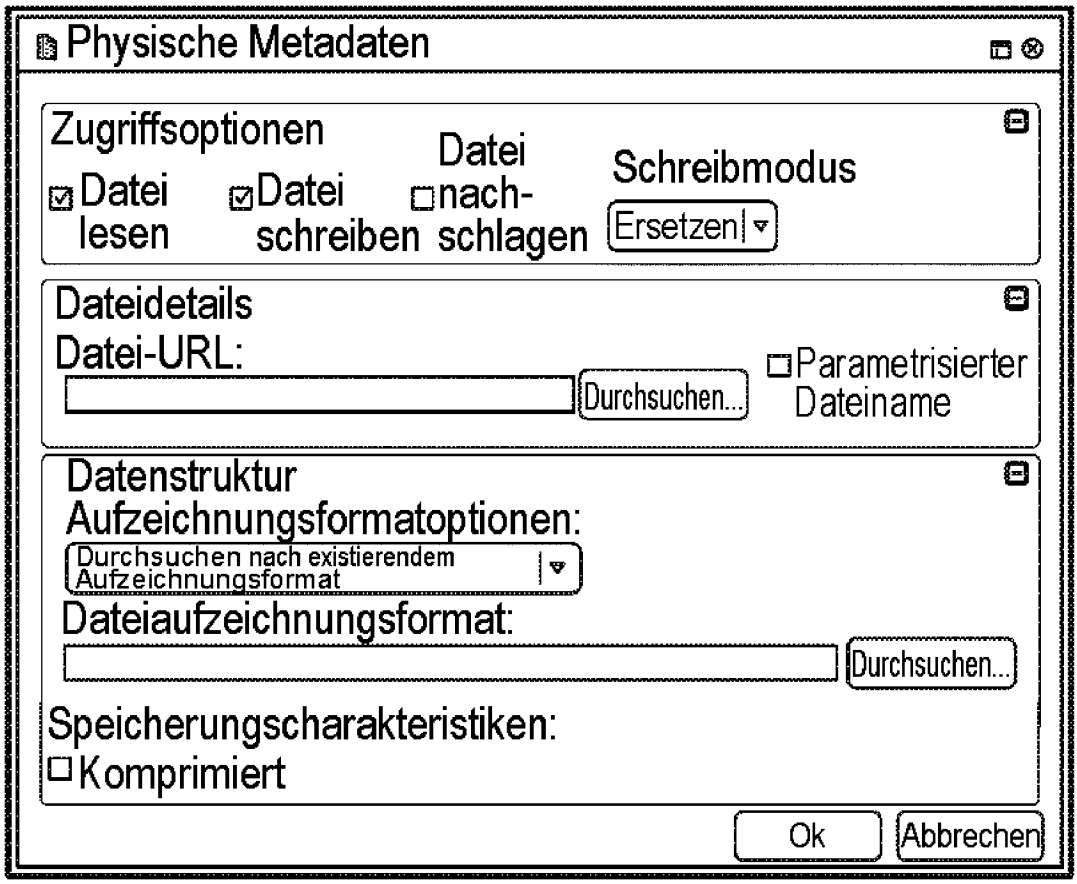

Physische Metadaten
Zugriffsoptionen
Datei lesen
Datei schreiben
Datei nach-schlagen
Schreibmodus
Ersetzen
Dateidetails
Datei-URL:
Durchsuchen...
Parametrisierter Dateiname
Datenstruktur
Aufzeichnungsformatoptionen:
Durchsuchen nach existierendem Aufzeichnungsformat
Dateiaufzeichnungsformat:
Durchsuchen...
Speicherungscharakteristiken:
Komprimiert
Ok
Abbrechen

FIG. 5C


Organisationsmetadaten
Optionale Metadaten:
Datensatzsystem:
Datensatzanwendung:
EME-Repositoriumswerte
Datei-RPfad-Name:
Anwenden
Voller-Datensatz-RPfad:
Ok
Abbrechen

FIG. 5D

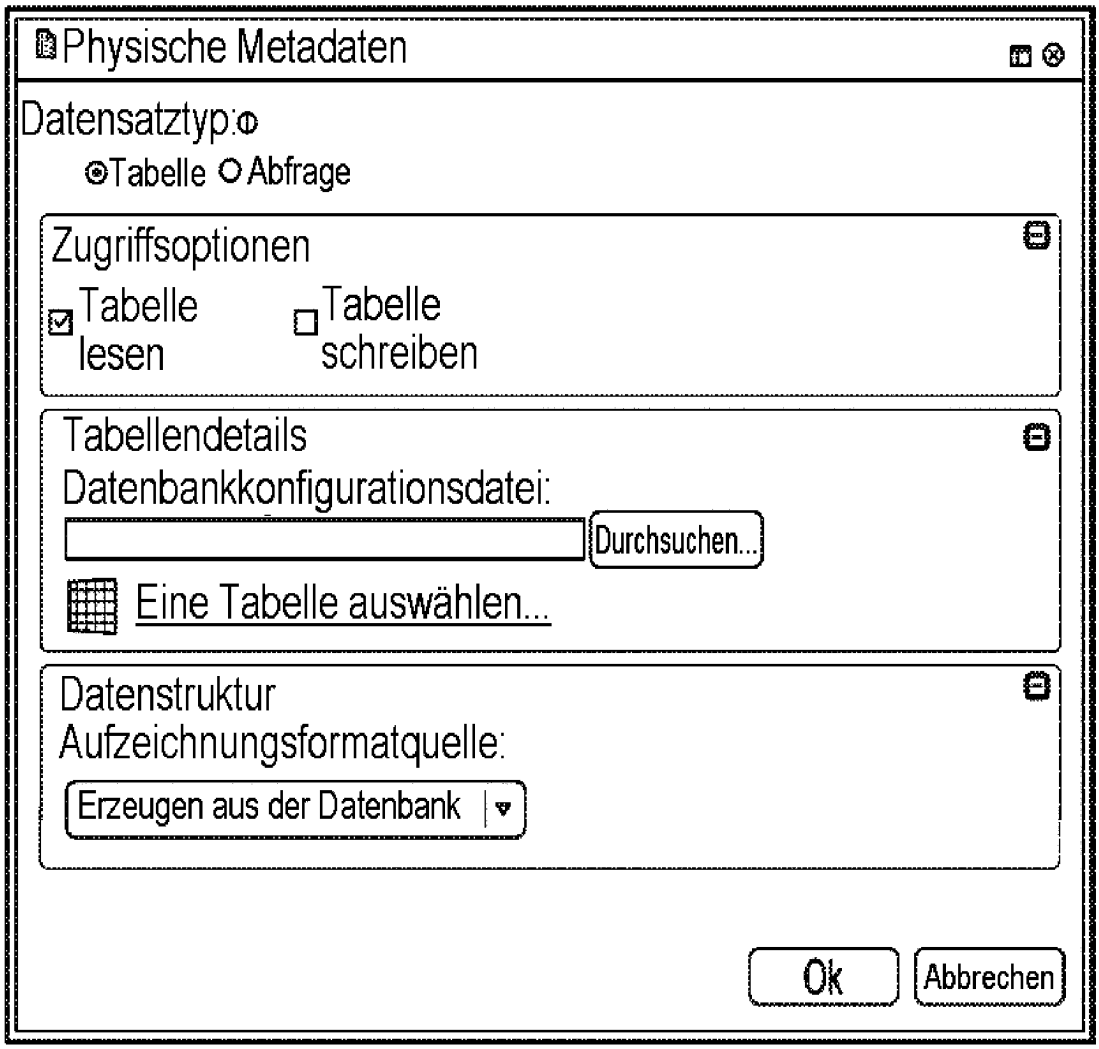

Physische Metadaten
Datensatztyp:
Tabelle
Abfrage
Zugriffsoptionen
Tabelle lesen
Tabelle schreiben
Tabellendetails
Datenbankkonfigurationsdatei:
Durchsuchen...
Eine Tabelle auswählen...
Datenstruktur
Aufzeichnungsformatquelle:
Erzeugen aus der Datenbank
Ok
Abbrechen

FIG. 5E

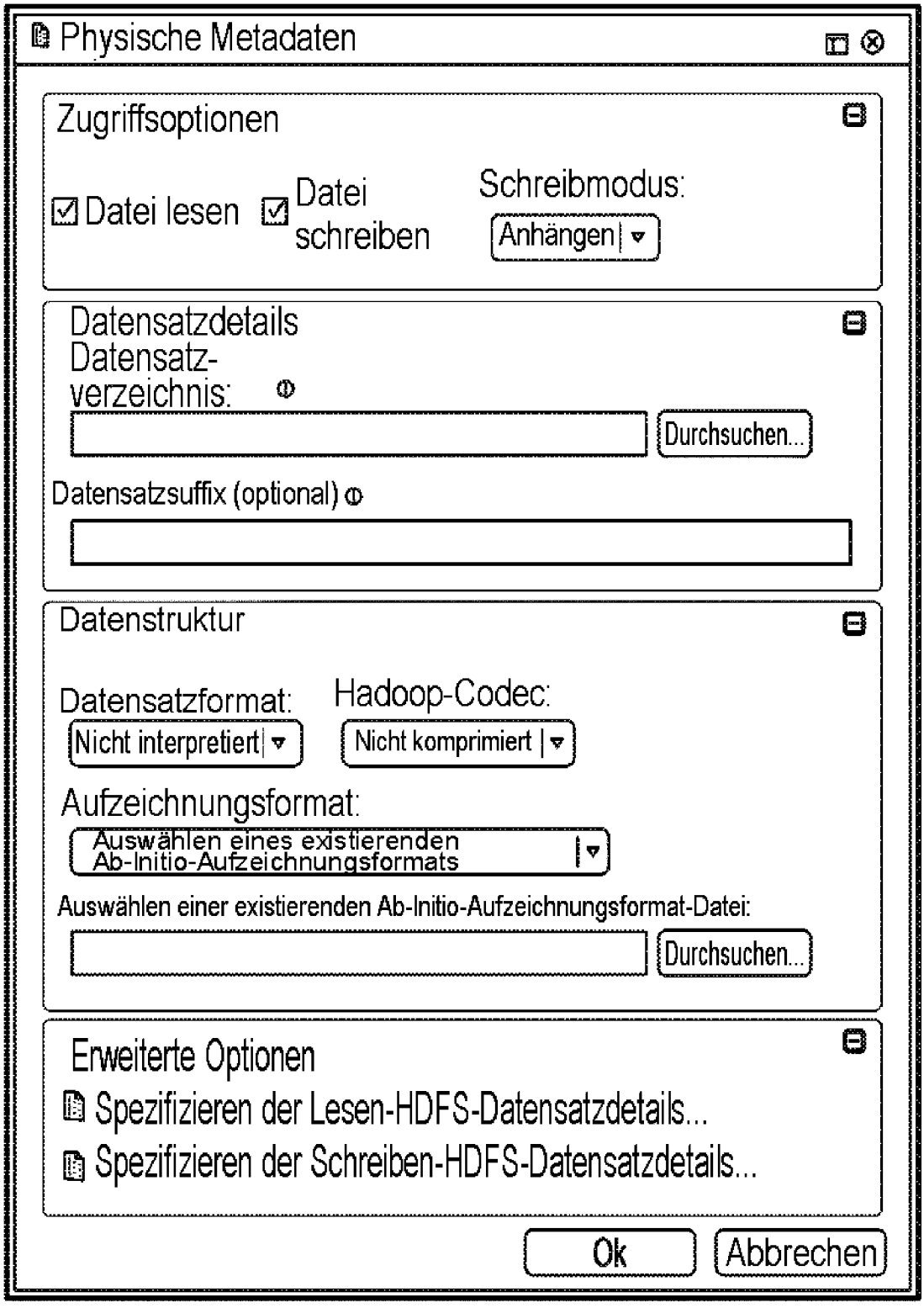

Physische Metadaten
Zugriffsoptionen
Datei lesen
Datei schreiben
Schreibmodus:
Anhängen
Datensatzdetails
Datensatz-verzeichnis:
Durchsuchen...
Datensatzsuffix (optional)
Datenstruktur
Datensatzformat:
Nicht interpretiert
Hadoop-Codec:
Nicht komprimiert
Aufzeichnungsformat:
Auswählen eines existierenden Ab-Initio-Aufzeichnungsformats
Auswählen einer existierenden Ab-Initio-Aufzeichnungsformat-Datei:
Durchsuchen...
Erweiterte Optionen
Spezifizieren der Lesen-HDFS-Datensatzdetails...
Spezifizieren der Schreiben-HDFS-Datensatzdetails...
Ok
Abbrechen

FIG. 5F

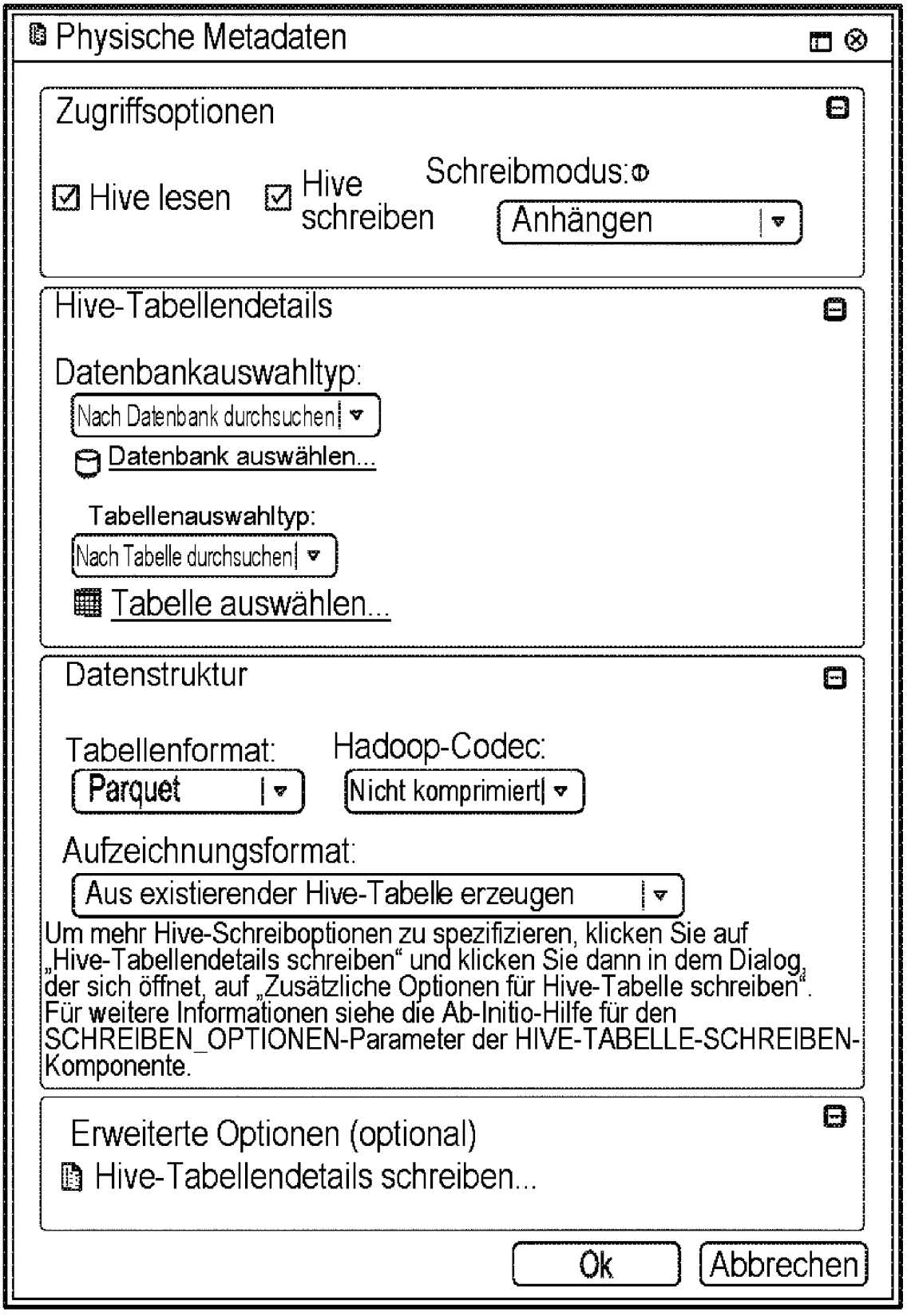

Physische Metadaten
Zugriffsoptionen
Hive lesen
Hive schreiben
Schreibmodus:
Anhängen
Hive-Tabellendetails
Datenbankauswahltyp:
Nach Datenbank durchsuchen
Datenbank auswählen...
Tabellenauswahltyp:
Nach Tabelle durchsuchen
Tabelle auswählen...
Datenstruktur
Tabellenformat:
Parquet
Hadoop-Codec:
Nicht komprimiert
Aufzeichnungsformat:
Aus existierender Hive-Tabelle erzeugen
Um mehr Hive-Schreiboptionen zu spezifizieren, klicken Sie auf „Hive-Tabellendetails schreiben“ und klicken Sie dann in dem Dialog, der sich öffnet, auf „Zusätzliche Optionen für Hive-Tabelle schreiben“. Für weitere Informationen siehe die Ab-Initio-Hilfe für den SCHREIBEN_OPTIONEN-Parameter der HIVE-TABELLE-SCHREIBEN-Komponente.
Erweiterte Optionen (optional)
Hive-Tabellendetails schreiben...
Ok
Abbrechen

FIG. 5G

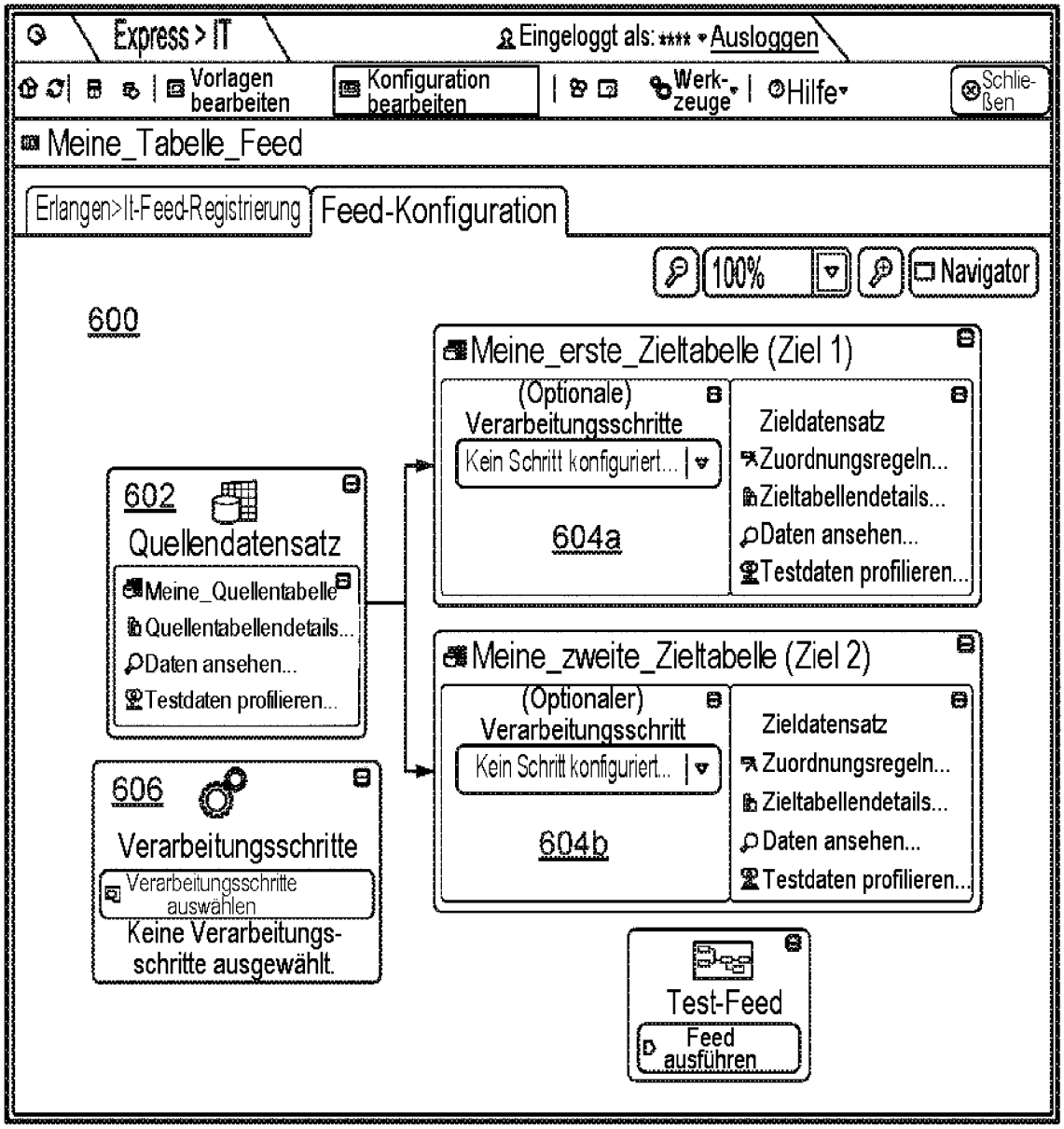

Express > IT
Eingeloggt als: **** Ausloggen
Vorlagen bearbeiten
Konfiguration bearbeiten
Werk-zeuge
Hilfe
Schlie-ßen
Meine_Tabelle_Feed
Erlangen>It-Feed-Registrierung
Feed-Konfiguration
100%
Navigator
600
Meine_erste_Zieltabelle (Ziel 1)
(Optionale) Verarbeitungsschritte
Kein Schritt konfiguriert...
604a
Zieldatensatz
Zuordnungsregeln...
Zieltabellendetails...
Daten ansehen...
Testdaten profilieren...
602
Quellendatensatz
Meine_Quellentabelle
Quellentabellendetails...
Daten ansehen...
Testdaten profilieren...
Meine_zweite_Zieltabelle (Ziel 2)
(Optionaler) Verarbeitungsschritt
Kein Schritt konfiguriert...
604b
Zieldatensatz
Zuordnungsregeln...
Zieltabellendetails...
Daten ansehen...
Testdaten profilieren...
606
Verarbeitungsschritte
Verarbeitungsschritte auswählen
Keine Verarbeitungs-schritte ausgewählt.
Test-Feed
Feed ausführen

FIG. 6

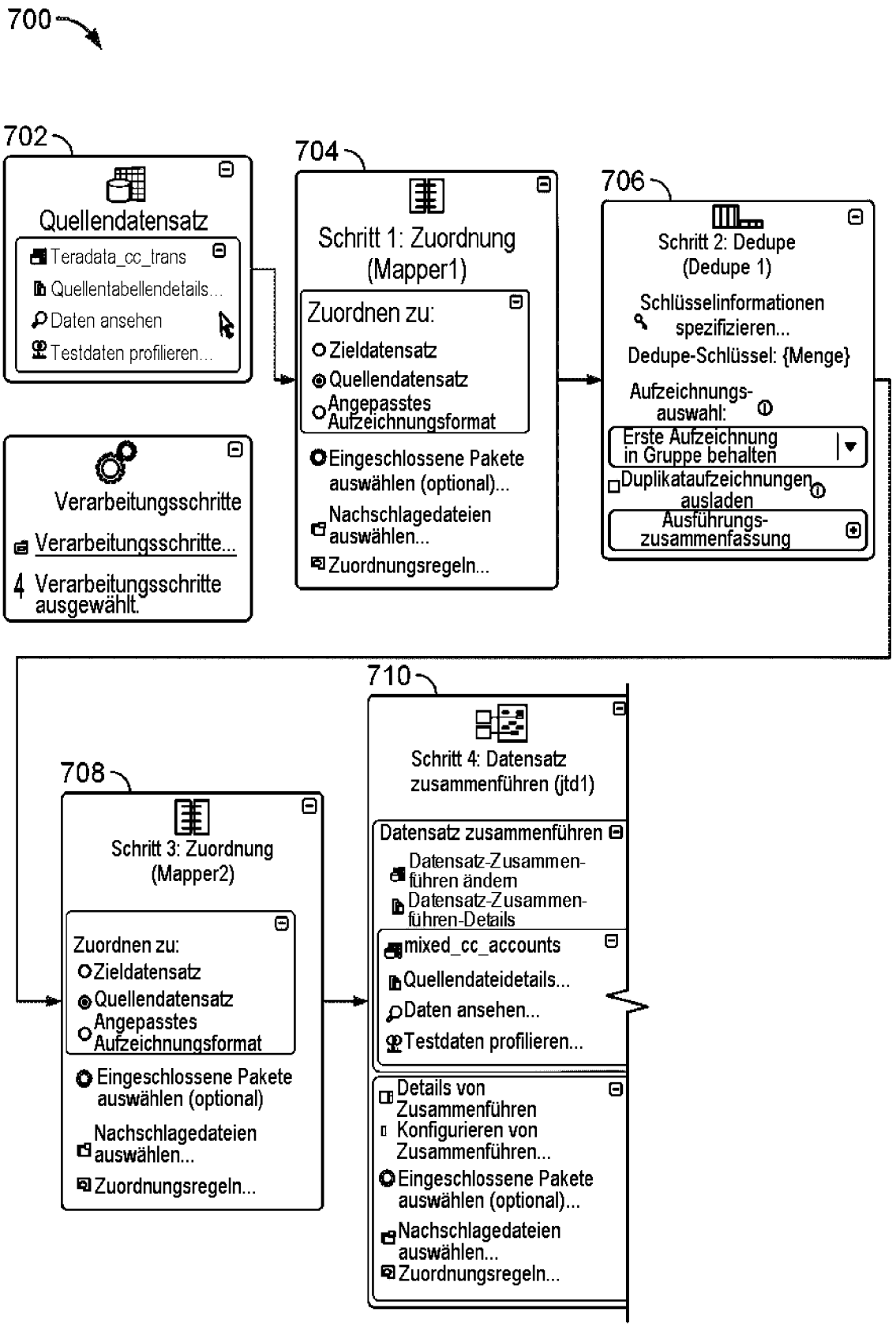

700
702
Quellendatensatz
Teradata_cc_trans
Quellentabellendetails...
Daten ansehen
Testdaten profilieren...
704
Schritt 1: Zuordnung (Mapper1)
Zuordnen zu:
Zieldatensatz
Quellendatensatz
Angepasstes Aufzeichnungsformat
Eingeschlossene Pakete auswählen (optional)...
Nachschlagedateien auswählen...
Zuordnungsregeln...
706
Schritt 2: Dedupe (Dedupe 1)
Schlüsselinformationen spezifizieren...
Dedupe-Schlüssel: {Menge}
Aufzeichnungs-auswahl:
Erste Aufzeichnung in Gruppe behalten
Duplikataufzeichnungen ausladen
Ausführungs-zusammenfassung
Verarbeitungsschritte
Verarbeitungsschritte...
4 Verarbeitungsschritte ausgewählt.
710
Schritt 4: Datensatz zusammenführen (jtd1)
Datensatz zusammenführen
Datensatz-Zusammen-führen ändern
Datensatz-Zusammen-führen-Details
mixed_cc_accounts
Quellendateidetails...
Daten ansehen...
Testdaten profilieren...
Details von Zusammenführen
Konfigurieren von Zusammenführen...
Eingeschlossene Pakete auswählen (optional)...
Nachschlagedateien auswählen...
Zuordnungsregeln...
708
Schritt 3: Zuordnung (Mapper2)
Zuordnen zu:
Zieldatensatz
Quellendatensatz
Angepasstes Aufzeichnungsformat
Eingeschlossene Pakete auswählen (optional)
Nachschlagedateien auswählen...
Zuordnungsregeln...

FIG. 7

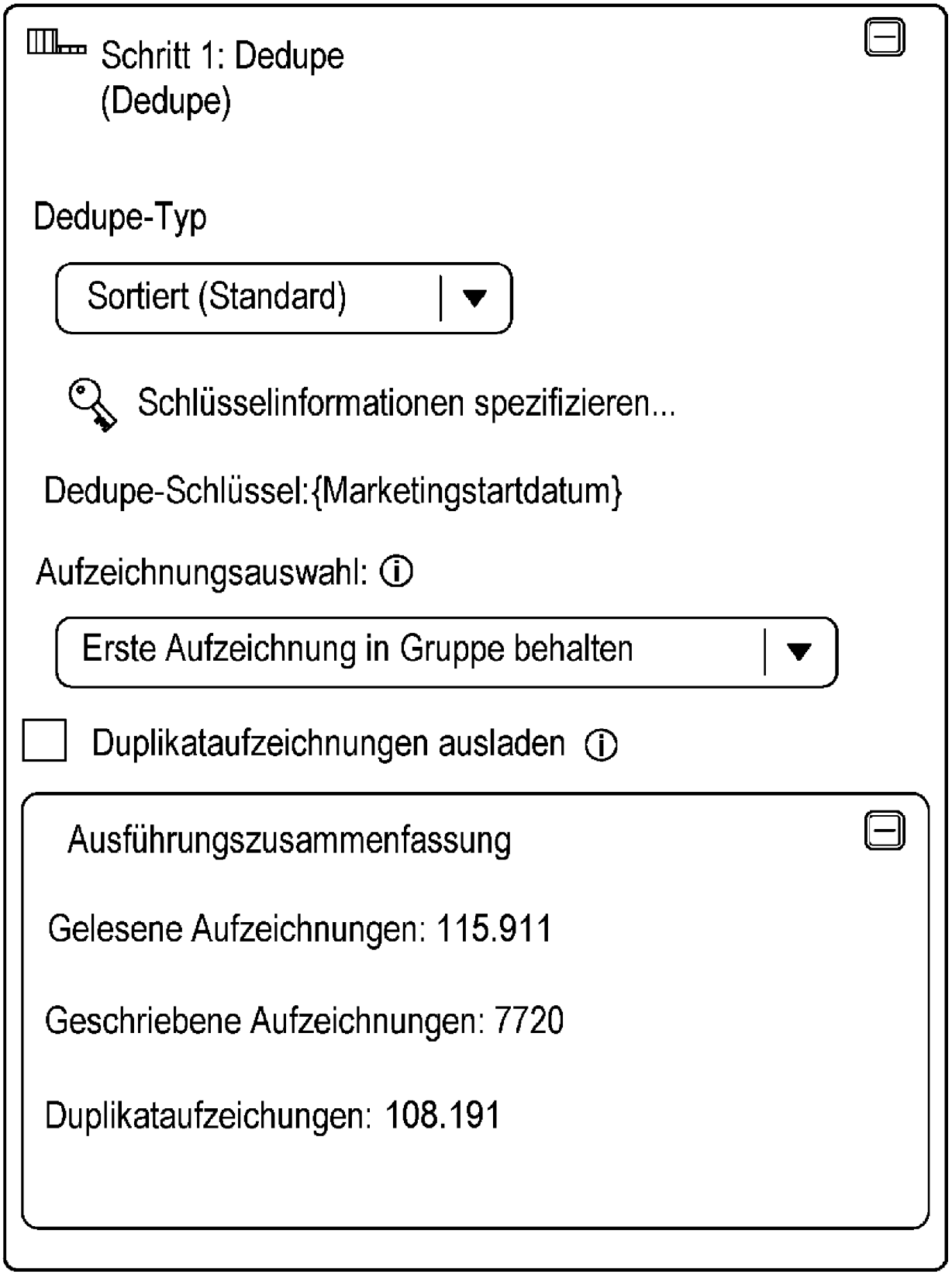

Schritt 1: Dedupe
(Dedupe)
Dedupe-Typ
Sortiert (Standard)
Schlüsselinformationen spezifizieren...
Dedupe-Schlüssel:{Marketingstartdatum}
Aufzeichnungsauswahl:
Erste Aufzeichnung in Gruppe behalten
Duplikataufzeichnungen ausladen
Ausführungszusammenfassung
Gelesene Aufzeichnungen: 115.911
Geschriebene Aufzeichnungen: 7720
Duplikataufzeichungen: 108.191

FIG. 8

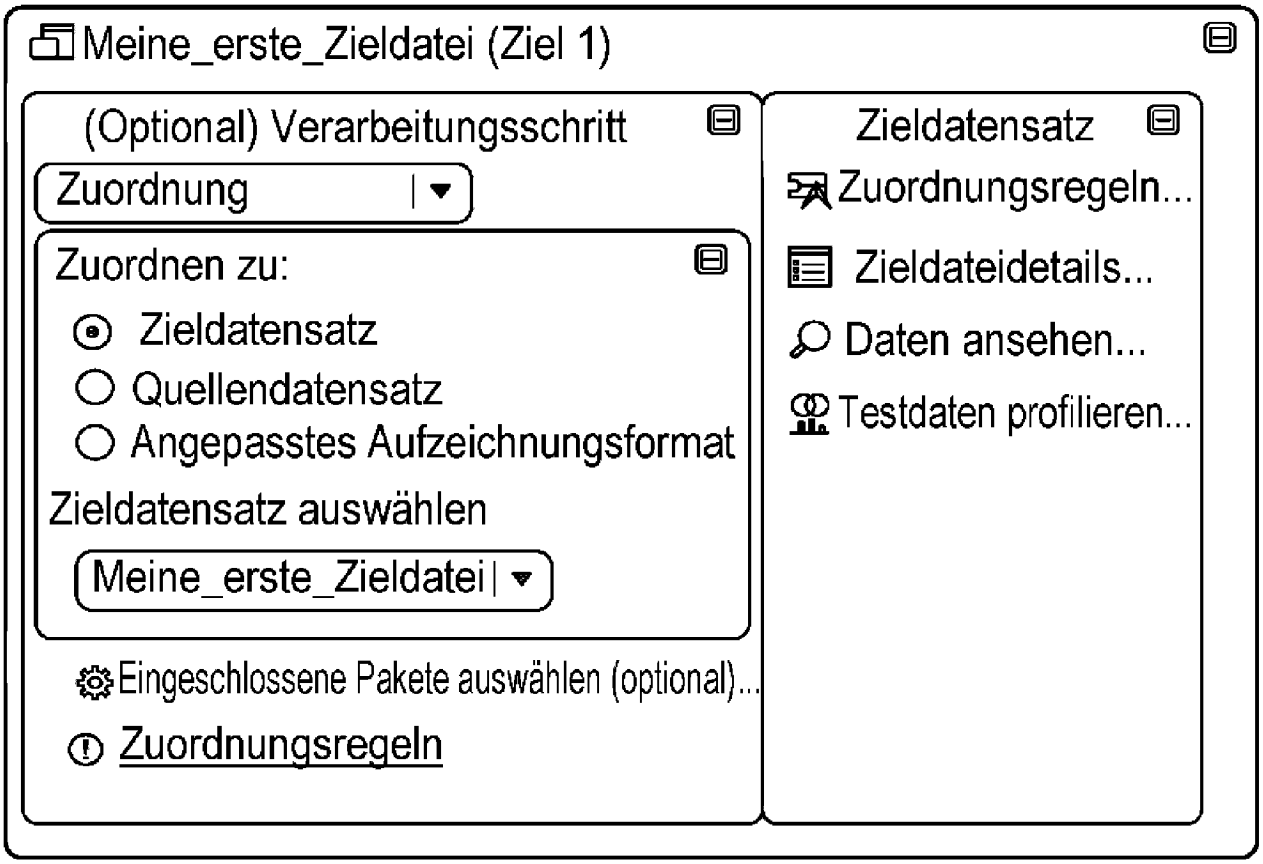

Meine_erste_Zieldatei (Ziel 1)
(Optional) Verarbeitungsschritt
Zuordnung
Zuordnen zu:
Zieldatensatz
Quellendatensatz
Angepasstes Aufzeichnungsformat
Zieldatensatz auswählen
Meine_erste_Zieldatei
Eingeschlossene Pakete auswählen (optional)...
Zuordnungsregeln
Zieldatensatz
Zuordnungsregeln...
Zieldateidetails...
Daten ansehen...
Testdaten profilieren...

FIG. 9A

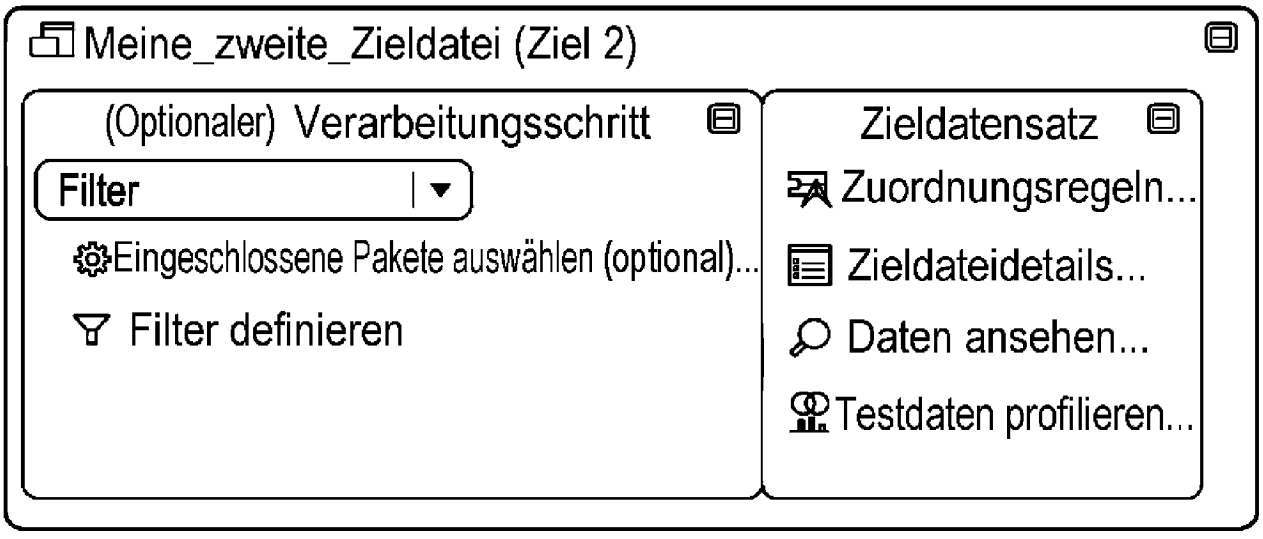

Meine_zweite_Zieldatei (Ziel 2)
(Optionaler) Verarbeitungsschritt
Filter
Eingeschlossene Pakete auswählen (optional)...
Filter definieren
Zieldatensatz
Zuordnungsregeln...
Zieldateidetails...
Daten ansehen...
Testdaten profilieren...

FIG. 9B

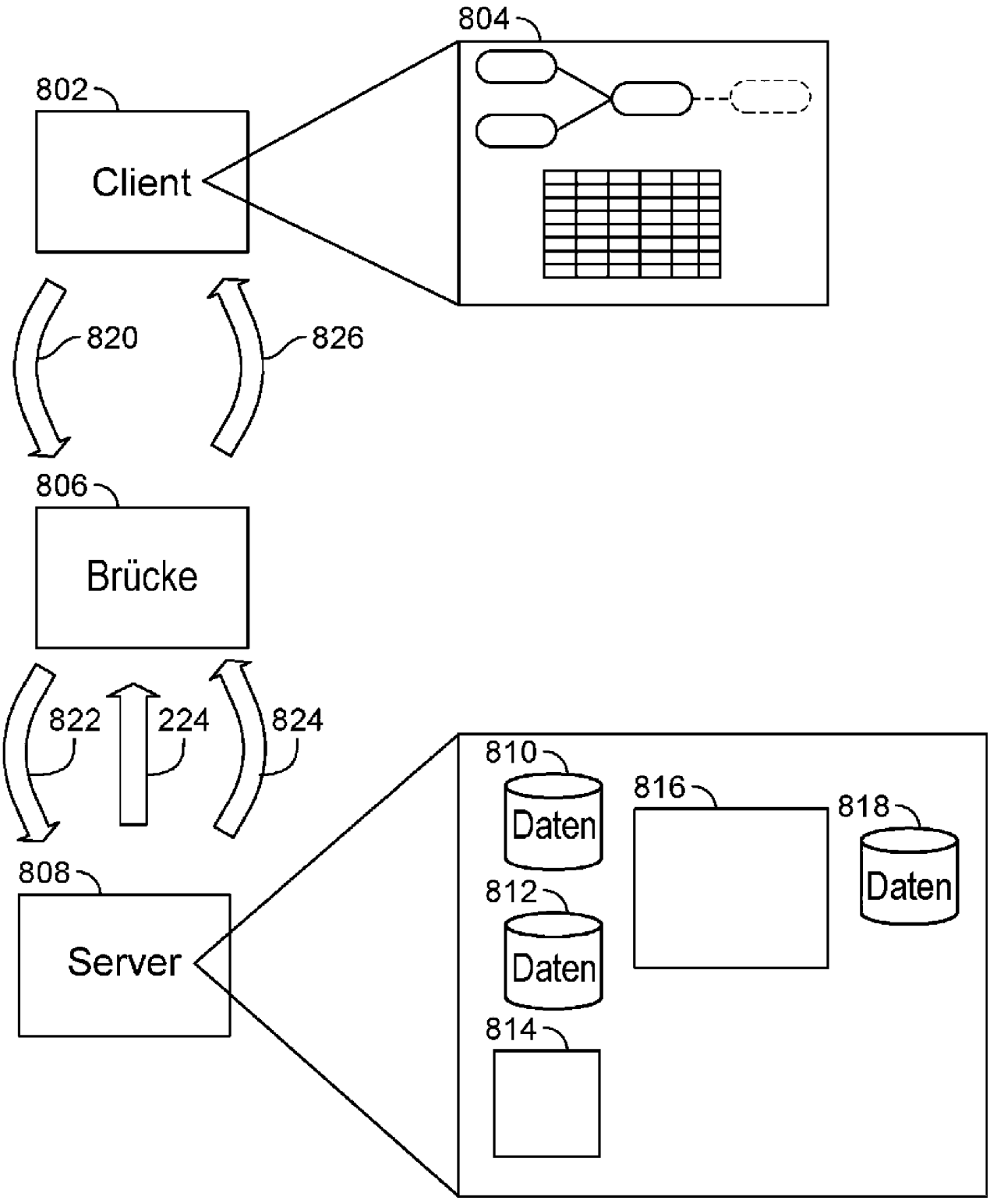

804
802
Client
820
826
806
Brücke
822
224
824
810
Daten
816
818
Daten
812
Daten
808
Server
814

FIG. 10