Historisierung und Analyse von Daten aus OEM Cloud Control in Hadoop

Ingo Reisky
OPITZ CONSULTING Deutschland GmbH
München

Matthias Fuchs
ISE Information Systems Engineering GmbH
Nürnberg

Schlüsselworte
Oracle Enterprise Manager (OEM), Metrikdaten, IT Analytics, Infrastruktur-Reporting, Oracle Business Intelligence Publisher (BIP), Apache Hadoop, Cloudera Data Platform (CDP), Apache Hive, Apache Flume, Apache Sqoop

Einleitung

Oracle Enterprise Manager 12c Cloud Control sammelt zahlreiche Daten auf fast allen Ebenen der IT-Infrastruktur. Gerade durch die Oracle Engineered Systems wird die Rolle des Cloud Control für das Management der verschiedenen Infrastrukturkomponenten immer wichtiger. Somit fallen nicht nur Datenbank-Metriken an, sondern auch Middleware- und Infrastruktur-Daten, wie z. B. von Loadbalancern.

Oft steht in der Praxis der Einsatz des OEM als Monitoring-Werkzeug im Vordergrund: Die Rohdaten (sog. "Current Metrics") der überwachten Ziele werden zeitnah, innerhalb weniger Tage nach Sammlung aggregiert und verlieren durch diese Verdichtung für spätere Datenanalysen an Informationsgehalt. Detaillierte Analysen der Infrastrukturdaten über längere Zeiträume sind deshalb nicht möglich.

Andererseits sprechen wirtschaftliche Gründe und Performancegründe gegen eine langfristige Speicherung von großen Datenmengen in der OEM Repository-Datenbank oder in anderen dedizierten RDBMS. In der Praxis beobachtet man zusätzliche Metrikdaten in der Größenordnung von 50-100 GB/Monat, was eine Realtime-Auswertung schwer macht. Damit bleiben aufgrund von Budgetrestriktionen Potenziale in den historischen Daten ungenutzt oder die Abfragegeschwindigkeit von Auswertungen geht in den zweistelligen Minutenbereich.

In diesem Vortrag wird gezeigt, wie Metrikdaten von z. B. Oracle Datenbanken und Oracle Fusion Middleware in einen Hadoop Cluster transferiert und dort gespeichert werden können, um sie anschließend mit Standard-Mitteln aus der Hadoop-Welt analysieren und visualisieren zu können. Dazu wurden Tests mit Oracle Enterprise Manager 12c Release 4 (Version 12.1.0.4) im Zusammenspiel mit Cloudera Data Platform 5 als Hadoop-Distribution durchgeführt.

Weiterhin werden die aktuellen Möglichkeiten im Zusammenspiel Oracle Datenbank und Hadoop, hier im speziellen mit Apache Hive und dem Hadoop Distributed File System (HDFS) als ausgewählte Bestandteile des Hadoop-Ökosystems betrachtet. Es werden sowohl Real Time-Ansätze angesprochen, als auch klassische Ladeszenarien aufgezeigt.

Problembeschreibung

Aus der Kernthese, dass in den mit Oracle Enterprise Manager 12c gesammelten Metriken der unterschiedlichsten Zieltypen in der Praxis oft ungenutztes Potential und Mehrwert für das Unternehmen schlummert, ergibt sich folgende zentrale Problemstellung:

Wie können die ohnehin für das Monitoring der Ziele mit OEM gesammelten Metrikdaten zusätzlich auch zu Analyse-, Planungs- und Reporting-Zwecken herangezogen werden? Und wie kann dies trotz bestehender Einschränkungen des vom Hersteller vorgegeben und auf schreibenden Zugriff ausgelegten Datenmodells der OEM Repository-Datenbank relativ performant und trotzdem für eine IT-Abteilung budgetfreundlich erreicht werden?

Datenhaltung

Diesem verborgenen Mehrwert der durch OEM gesammelten Metrikdaten stehen in der Praxis mehrere Faktoren entgegen: Die Einschränkungen im Datenmodell resultieren aus einer Optimierung der Datenbank auf effektives Schreiben der vom Oracle Management Server (OMS) von den Agenten empfangenen Metrikdaten in die Repository-Datenbank. Ein späteres Aggregieren und zeitgesteuertes Löschen von Detaildaten ist notwendig um die Performance der Anwendung zu gewährleisten.

Es können diverse Änderungen am Repository des Cloud Controls durchgeführt werden wie z.B. eine Erhöhung der Zeit der Datenvorhaltung. Die dadurch erhöhten Kosten für die Datenhaltung dieser großer Mengen von nicht-fachlichen Daten in relationalen Datenbanksystemen „nur“ zu Analyse-, Planungs- und Reportingzwecken sind in der Praxis aufgrund von Budgetrestriktionen schwer zu begründen (unabhängig davon, ob es um die OEM Repository-Datenbank oder andere hoch-performante RDBMS/DWH im Unternehmen geht). In der Praxis können Metrikdaten in einer Größenordnung von beispielsweise 50-100 GB zusätzlich je Monat produziert werden. Je nach Zeithorizont der Analyseanforderungen (=Mindest-Aufbewahrungsdauer) lässt sich also tatsächlich von großen Datenmengen sprechen, die in den Terrabyte Bereich gehen können. Um bei einem OLTP System eine hohe Performance in diesen Größenordnungen zu gewährleisten sind hohe Investitionen nötig.

Reporting

Weiterhin ist bezüglich der Nutzung der OEM-eigenen Reportingwerkzeuge Oracle Business Intelligence Publisher (BIP) und Information Publisher in diesem Kontext zu beachten: BIP hat sich in der Praxis bewährt für Reports mit begrenzter Komplexität der Abfragen, aber komplexe Abfragen der Metrikdaten (z.B. mit Pivot-Funktion in SQL Statements) dauern zu lange für interaktives Online-Reporting oder sind während der Hauptlastzeiten mit Rücksicht auf die Repository-Datenbank gar nicht zu empfehlen. Im Gegensatz dazu ist der Information Publisher seit längerem als Komponente im OEM abgekündigt und aus diesem Grund für ein Reporting nach der alten bisherigen Art und Weise nicht zu empfehlen. Aber unabhängig davon, mit welchem OEM-integrierten Werkzeug (Business Intelligence Publisher, Information Publisher oder SQL*plus) gearbeitet wird, die zuvor ausgeführten Einschränkungen des Repository-Datenmodells in diesem Kontext gelten auch hierbei (Zeit der Vorhaltung).

Daher die Motivation für den folgenden Lösungsansatz, der sowohl die Einschränkungen im Datenmodell der OEM Repository-Datenbank umgeht, größere Datenvolumina relativ kostengünstig über längere Zeiträume speichern lässt als auch weitere Zusatznutzen bietet.

Lösung Big Data

Unter den ausgeführten Prämissen (OEM Repository ist auf Monitoring spezialisiert, Kosten der Datenhaltung sollen begrenzt werden, etc.) haben wir unterschiedliche Möglichkeiten im Oracle Enterprise Manager 12c untersucht, Metrikdaten und zusätzliche Daten aus dem OEM Gesamtsystem heraus zu transferieren in ein separates Reporting-Analyse-System, das kostengünstiger als dedizierte relationale Datenbanken die langfristige Speicherung der Metrikdaten erlauben und kostengünstige, aber bewährte Reporting- und Analysewerkzeuge für große Datenmengen bieten sollte.

Dabei fiel die Wahl auf das Apache Hadoop Ökosystem von Open-Source Datenspeicherungs-, Aggregations- und Auswertungswerkzeugen, da die Verwendung von Commodity Hardware als Grundidee von Hadoop zur verteilten, ausfalltoleranten Speicherung und Verarbeitung von großen Datenmengen tendenziell eine kostengünstigere massenhafte Datenhaltung als in spezialisierten, auf Geschwindigkeitsoptimierung getrimmten RDBMS/DWH erlauben kann. Wir haben konkret mit der Hadoop-Distribution von Cloudera getestet ("Cloudera Data Platform", CDP Version 5) aufgrund der Partnerschaft zwischen Oracle und Cloudera, die beispielsweise in der Verwendung von CDP im Oracle Engineered System "Oracle Big Data Appliance" Ausdruck findet. Ein analoges Vorgehen wäre auch mit anderen Hadoop-Distributionen (z.B. der Distribution von Hortonworks) möglich.

Übernahme der Datenbank-Metrikdaten in Hive

Apache Hive ist eine relationale Datenbank unter Hadoop und auch als Komponente in der Hadoop-Distribution von Cloudera vorhanden. Ähnlich einem klassischen Data Warehouse-Ansatz werden die Daten aus dem Life-System, in unserem Fall der OEM Repository-Datenbank, in die Cloudera Data Platform importiert. Dieser Prozess kann Realtime aufgesetzt werden (beispielsweise mit Oracle Golden Gate), oder wiederholend im Batch-Verfahren: Im Vortrag werden wir detailliert auf Möglichkeiten des Ladens mit Open Source Tools wie Apache Scoop eingehen. Prinzipiell wäre dies aber auch mit Oracle Data Integrator (ODI) möglich.

Es sollen nur eine Übernahme der Metrikdaten und Konfigurationsdaten der Zielsysteme durchgeführt werden. Eine Transformation oder Optimierung dieser Daten soll im ersten Schritt nicht erfolgen. Es wird also eine RAW Staging-Schicht aufgebaut. Nachgelagert werden kurz Möglichkeiten der Analyse auf Basis des Hadoop Ökosystems aufgezeigt.

Gezielte Übernahme direkt aus dem Cloud Control

Neben dem direkten Auslesen der OEM Repository-Datenbank bietet Cloud Control auch die Möglichkeit, Daten über Metrikerweiterungen (sog „Metric Extensions“) oder per „Data Exchange Connector“ auszutauschen. Hier werden beide Wege aufgezeigt und Einsatzszenarien beschrieben.

Hintergrund: Metric Extensions (ME; früher „User-Defined Metrics“ genannt) erweitern die Standard-metriken im OEM um individuelle Metriken und basieren auf selbst-erstellten, zentral im OEM gepflegten Skripten (Shell, SQL, WLST, etc.) mit dezentraler, lokaler Ausführung durch die Agenten auf den Zielsystemen. Agenten können über ein entsprechend angepasstes Metric Extension-Skript den Hadoop Cluster mit wählbarer Frequenz ansprechen und die Daten direkt dorthin transferieren, beispielsweise an Apache Flume. Bei Bedarf kann zusätzlich die sog. Historisierung der Metric Extension deaktiviert werden, um Datenvolumen im OEM Repository zu sparen.

Und was geht noch in Hadoop?

Zusätzliche Vorteile bieten sich, sobald die Metrikdaten erfolgreich in den Hadoop Cluster übertragen werden konnten:

Durch die Verwendung des Hadoop Distributed File System (HDFS) ist es nicht nur möglich strukturierte Daten abzuspeichern, sondern auch unstrukturierte Daten wie z.B. Logfiles. Somit kann man mit Hadoop-Werkzeugen wie Apache Flume auch ein Streaming von Logfile-Daten realisieren. Es kann aber stattdessen auch nur ein einfaches Kopieren der Logdateien in das HDFS stattfinden. Mit weiteren Tools ist anschließend im Hadoop Cluster eine Mustererkennung möglich.

Außerdem sind Metrik-Rohdaten im höchsten im Cloud Control existierenden Detailgrad („Current Metrics“) read-only / unveränderlich in Hadoop speicherbar, unabhängig von den im OEM konfigurierten Aggregations- bzw. Purging-Intervallen für Metrikdaten, also konkret bei Bedarf länger als ein Jahr (Standardwert der gröbsten Aggregationsstufe) feingranular nutzbar.

Sind diese Daten in Hive oder HDFS persistiert, können umgekehrt mit Oracle Big Data SQL gezielt ausgewählte Datensätze aus der großen Menge an historischen Rohdaten in einer relationalen Datenbank zur weiteren Auswertung mit den Standardwerkzeugen des Unternehmens zur Verfügung gestellt werden.

Zusammenfassung und Ausblick

Zusammenfassend lässt sich festhalten, dass der gewählte Lösungsansatz mit Transfer der Metrikdaten aus dem Oracle Enterprise Manager in die Hadoop-Welt eine Möglichkeit zur Umgehung der dargestellten Einschränkungen des OEM Datenmodells. Mit der dauerhaften, langfristigen Speicherung der Rohdaten der Ziel-Metriken in Hadoop bieten sich alle Vorteile, mit den gängigen Werkzeugen des Hadoop-Ökosystems unterschiedliche Auswertungen und Aggregationsstufen auf denselben Rohdaten zu fahren, mit maßgeschneiderten Abfragetools/Datenmodellen je nach Auswertungsanforderung, bis hin zu Online IT-Analytics und Visualisierung der Metriken. Insgesamt ergibt sich mit der Persistierung der Metrikdaten in Hadoop eine Lösung, die viele Teilaspekte und Anforderungen abdecken kann.

Nachdem zukünftig nicht nur Daten aus dem Oracle Enterprise Manager Cloud Control für eine Analyse der unternehmensweiten IT-Infrastruktur notwendig sein werden, können durch den offenen Ansatz von Hadoop weitere, heterogene Quellen angebunden und zusammengeführt werden. Hierzu gehören beispielsweise Daten aus dedizierten Monitoringsystemen der Betriebssystem- oder Netzwerkschicht sowie Daten aus dem Performance Management und End-to-End-Monitoring. Auch eine Anreicherung mit detaillierten, aus Verknüpfungen innerhalb einer SQL Datenbank gewonnenen Daten ist möglich (z.B. erweiterte Daten, die von Cloud Control nicht erfasst werden). Somit verlässt man schnell den Bereich des reinen Monitorings hin zu einer Governance oder Ressourcenüberwachung innerhalb eines Unternehmens. Ebenso können Daten von einer Applikation, die sich in der Cloud befindet, eingebunden werden.

Kontaktadressen:

Ingo Reisky
OPITZ CONSULTING Deutschland GmbH
Standort München
Weltenburger Str. 4
D-81677 München

Telefon: +49 (0) 89-680098-1489
Fax: +49 (0) 89-680098-4400
E-Mail ingo.reisky@opitz-consulting.com
Internet: http://www.opitz-consulting.com

Matthias Fuchs
ISE Information Systems Engineering GmbH
Südwestpark 70
D-90449 Nürnberg

Telefon: +49 (0) 172-8288751
E-Mail matthias.fuchs@ise-informatik.de
Internet: www.ise-informatik.de