# Big Data Preparation Cloud Service
## ETL und DQ und noch mehr

**Dr. Holger Dresing**
**Oracle Deutschland B.V. & Co. KG**
**Hannover**

**Schlüsselworte**
Data Preparation, Big Data, Cloud, Datenintegration, ETL, Data Quality, PaaS, Platform as a Service

### 1. Einleitung

Im Big Data Umfeld wird von großen aber auch von unstrukturierten Datenmengen gesprochen. Um aus den Datenmengen aussagefähige und geschäftsrelevante Informationen zu gewinnen, müssen die Daten aufbereitet und in auswertbare Strukturen überführt werden. Das Aufbereiten dieser Daten und die semantische Analyse wird aktuell als ein großes Problem im Big Data Umfeld gesehen, bis zu 90% der Auswände fließen dort hinein.
Oracle Big Data Preparation (kurz: BDP) Cloud Service vereinfacht es dem Anwender, mit diesen Daten zu arbeiten: Die Datenaufbereitung wird automatisiert, auf ihren Inhalt geprüft, ergänzt und korrigiert ohne manuelle Eingriffe. BDP steht als skalierbare Cloud Anwendung zur Verfügung, die auf Hadoop/Spark mit natürlicher Sprachverarbeitung und Referenz Datensätzen basiert. Dafür wird auf eine für maschinelles Lernen entwickelte Engine zurückgegriffen, die das Vorbereiten und Bearbeiten der Daten unterstützt. Wiederkehrende Geschäftsstrukturen können automatisiert verarbeitete werden und im Fehlerfall werden standardisierte Verfahren eingefügt. Der Anwender möchte sich auf die geschäftlichen Zusammenhänge in den Daten konzentrieren, daher werden klassische Data Quality Verfahren wie Profiling, Cleansing, Standardisierung und Ergänzung mit integriert. Zusammenfassend wird die Zeit für die Aufbereitung erheblich verkürzt und somit die Kosten und Zeitaufwände gegenüber traditionellen ETL- und Data Quality-Verfahren erheblich verkürzt.

### 2. Möglichkeiten von Big Data Preparation

Big Data Preparation Cloud Services bietet verschiedene Möglichkeiten, um Daten aufzubereiten, eine semantische Analyse der Daten vorzunehmen und die Vorgänge zu standardisieren.

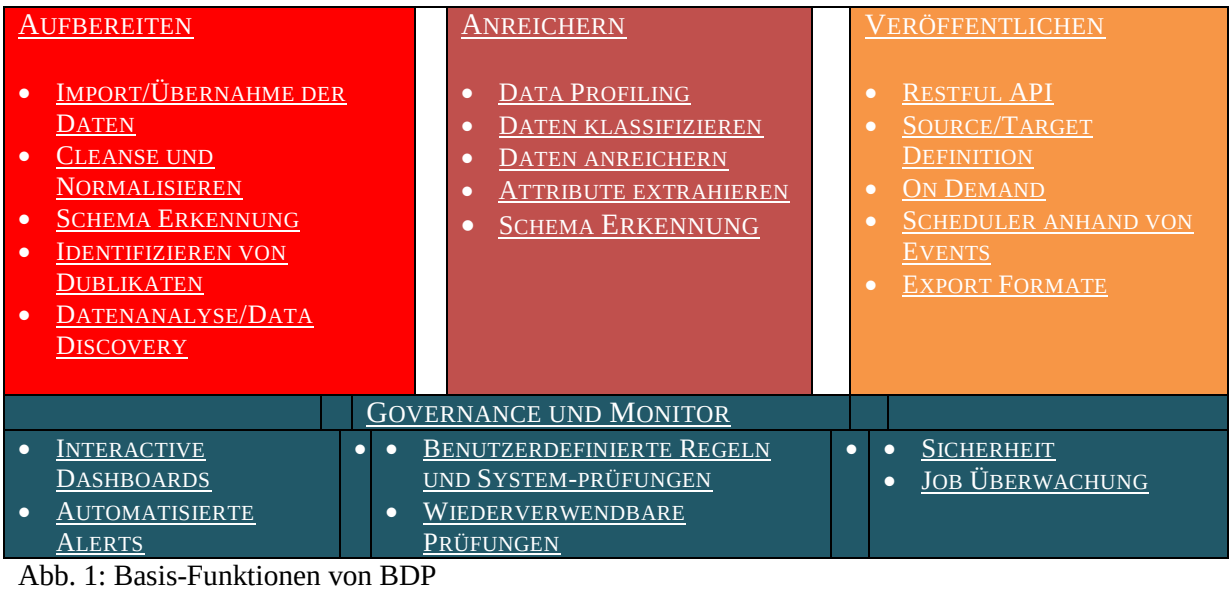

Abb. 1: Basis-Funktionen von BDP

***Daten aufbereiten und anpassen***
Zu diesem Bereich gehört:

- *Statistisches Profiling* – statistische Analyse numerische Daten sowie Häufigkeiten von Ausdrücken in Texten
- *Cleansing, Normalisierung* – nicht relevante Zeichen werden entfernt, Standardisierung von Inhalten wie z. B. Datumswerte
- *Daten Reparatur* – Erkennen und Bereinigen von möglichen Inkonsistenzen in den Daten
- *Data-Enrichment* – semantische Analyse und ergänzen von relevanten Daten
- *Explizite Schema Erkennung* – Identifizieren von Schemata/Metadaten, die explizit in Dateiköpfen, Feldern oder anderen Informationen enthalten sind
- *Identifizieren von Duplikaten* – Erkennen von Duplikaten in den Daten

***Semantische Metadaten Erkennung, Anreicherung und Korrelation***
Zu diesem Bereich gehört:

*Klassifikation und Attribute Extraktion* – identifiziert Kategorien in den Daten und Charakteristika wie Attribute, Eigenschaften oder Schemata

*Implizite Schema Erkennung* – häufig können Schemata anhand ihrer Dateninhalte erkannt werden, z. B. bei Email Adressen, postalischen Adressen oder Namen. Dieser Service erlaubt es, viele typische Attribute oder Eigenschaften in Schemata automatisch zu erkennen

***Monitoring und Governance***
Zu diesem Bereich gehört:

*Dashboard* – alle in BDP verarbeiteten Prozesse und Datensätze können über ein Dashboard überwacht und analysiert werden.

*Email-Benachrichtigungen* – Emails benachrichtigen über die Ausführung von Jobs und informieren über den Status, Warnungen oder Fehler

***Publishing/Veröffentlichung***
Zu diesem Bereich gehört:

*Quellen/Ziele* – BDP unterstützt viele unterschiedliche Quellen und Ziele wie Oracle Storage Cloud, andere Cloud Stores sowie URL Quellen.

*Ausführungsmethoden* – Prozesse können erstellt und interaktiv gestartet werden über das Benutzerinterface. Ebenso können Prozesse automatisch per Scheduler oder bei Eintreffen neuer Daten automatisch gestartet werden.

*Formate* – BDP unterstützt den Export von Daten über viele Standardformate wie CSV, XLS, XML, JSON, um damit typische on-premise BI-, Analyse- oder ETL-Prozesse mit Daten zu versorgen.

## 3. Arbeiten mit BDP an einem Beispiel

Um mit BDP arbeiten zu können, wird ein Oracle Storage Cloud Service und ein Oracle Big Data Preparation Cloud Service benötigt.

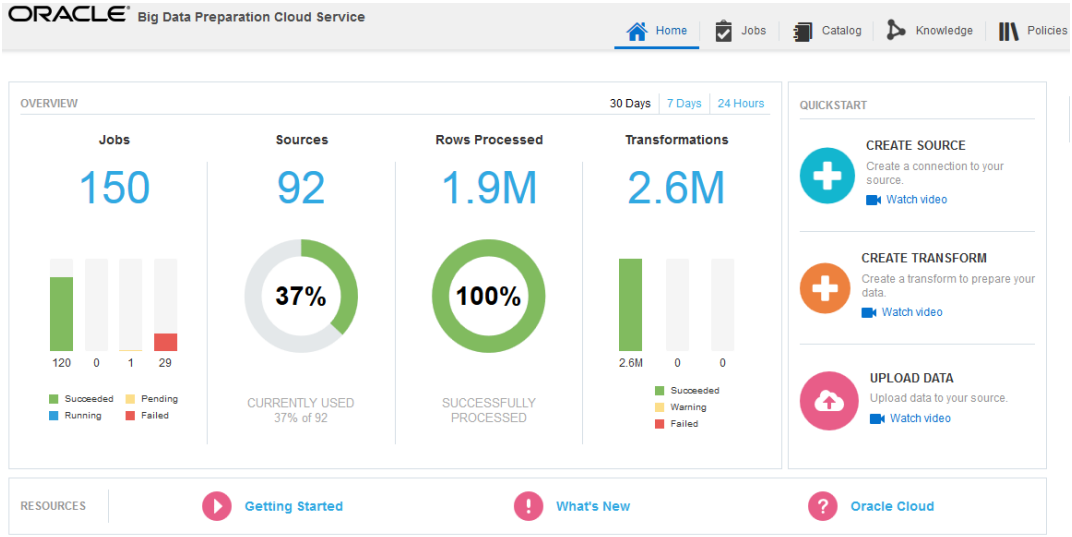


Abb. 2: Dashboard von BDP

Nachdem der Benutzer sich an BDP angemeldet hat, muss er zunächst eine Source erstellen (siehe Abb. 2). Jede Source benötigt eine eindeutige Bezeichnung. Es werden Text-Dateien, CSV, XML, Json, Excel und Word-Dateien ebenso komprimierte Formate wie ZIP, GZIP oder Tar unterstützt. Anschließend können die Daten hochgeladen werden (Upload Data). Dann kommt der Transform Schritt. Ein „transform“ ist eine Gruppe von Reparatur- und Ergänzungsregeln, die auch die Datensätze angewandt werden. Nachdem die Source hochgeladen wurden, wird dieser Schritt automatisch angestoßen und der Benutzer wird auf die „main authoring page“ geführt. Zunächst benötigt der Transformationsvorgang noch einen Namen und es wird nach führenden Kopfzeilen gefragt, wie das bei der Verarbeitung von Dateien üblich ist.

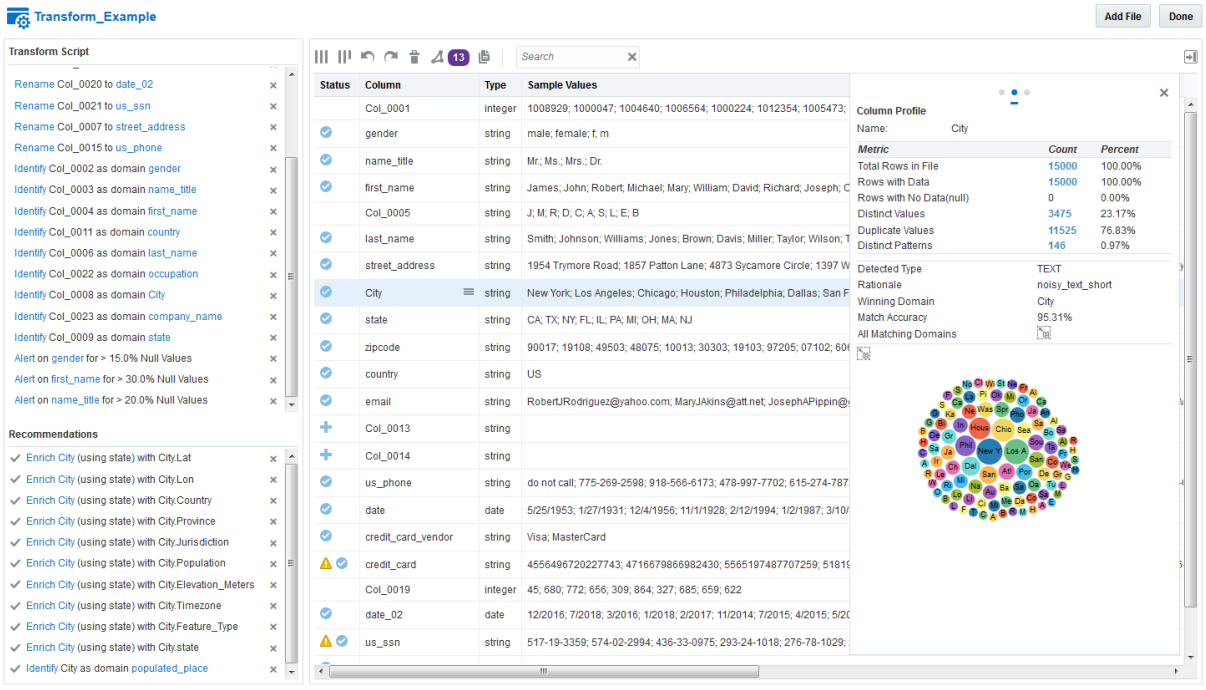


Abb. 3: Main Authoring Page in BDP

In Abb. 3 ist in der Mitte die Interpretation der einzelnen Spalten zu sehen. Im der Beispieldatei waren zwar Spaltenköpfe vorhanden, aber die erste Spalte wurde übersprungen und die automatische Analyse wurde aktiviert. Man kann mit dem „Metadata View Icon sich die Daten im Detail ansehen. Im linken oberen Bereich wird das aktuelle „Transform-Skript“ angezeigt und im linken unteren Bereich mögliche Bemerkungen für das „Transform-Skript“. In der rechten Spalte findet man eine Übersicht, wie sie für Profiling Werkzeuge üblich ist. Abhängig vom Datentyp und der Anzahl gefundener Werte kann sich die Darstellung ändern.

In Abb. 3 geht es um die Spalte „City“, hier wird das Vorkommen von Daten angezeigt mit Angaben wie z. B. Anzahl Zeilen, Anzahl Zeilen mit Daten, Anzahl unterschiedlicher und gleicher Werte. Darunter wird graphisch angezeigt, wie oft die Werte gefunden wurden. Betrachtet man den Wert „Match Accurcy“ so heißt dass, 95,31% der Citys sind dem semantische Analysesystem von BDP bekannt. Los Angeles, New York, Chicago und Houston sind häufig vorkommende Städte.
Abb. 4 zeigt ein Beispiel für den Datentyp „Datum“.

| Metric | Count | Percent |
|---|---|---|
| Total Rows | 15000 | 100.00% |
| Empty Rows | 0 | 0.00% |
| Total Columns | 23 | 100.00% |
| Empty Columns | 2 | 8.70% |
| Filtered Columns | 0 | 0.00% |
| Singleton Columns | 1 | 4.35% |
| Processed Columns | 23 | 100.00% |
| Classified Columns | 19 | 82.61% |
| Cols with Recommendations | 13 | - |
| Total Recommendations | 95 | - |

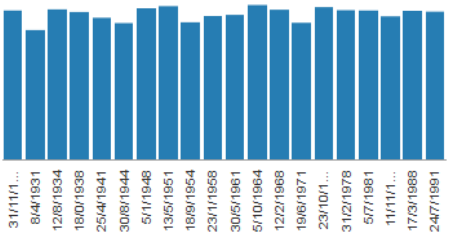


Abb. 4: Profiling über eine Datumsspalte

Für Spalten kann es sogenannte „Recommandations“/Bemerkungen geben.

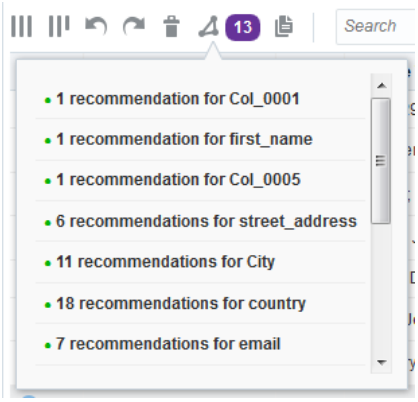


Abb. 5: Beispiele für Anmerkungen

Folgende Bemerkungen von BDP generiert:

- Vor den Spalten „credit_card“ und „us_ssn“ steht ein Icon mit einem Aufrufungszeichen: Kreditkartennummern oder Sozialversicherungsnummern sind besonders schützenwerte Daten, die von BDP automatisch erkannt werden und können als besonders schützenswerte (obfuscate) markiert werden. Es gibt weitere Vorschläge zur Anzeige solcher Spalten. In diesem Fall wird u. a. vorgeschlagen, nur die letzten 4 Zeichen anzuzeigen.
- In der ersten Spalte „col_0001“ ist jeder Wert eindeutig. Daher wird diese Spalte als Primärschlüssel benutzt. Dazu wird die Spalte in „customerID“ umbenannt.
- In der Spalte „Gender“ (Geschlecht) kommen vier Werte vor. Werte können durch andere Werte ersetzt werden. In diesem Fall wird „m“ durch „male“ und „f“ durch „female“ ersetzt.

- Zwischen den Spalten „first_name“ und „last_name“ wurde eine zusätzliche Spalte mit einzelnen Buchstaben gefunden. Für diese wurde „middle_name“ als Bezeichnung vorgeschlagen.
- Es wurden weitere Spalten basierend auf der Spalte „City“ eingefügt: Anzahl Einwohner und Gerichtsstand (jurisdiction). Dies sind semantische Ergänzungen, die BDP bereithält.
- Die Spalte „Country“ wurde gelöscht, da dort immer der Wert „US“ enthalten war.
- In der Spalte „email“ wird ein @ erwartet. Vorher steht der Benutzer und danach die Email-Domäne. Mit regulären Ausdrücken kann man daraus diese zwei Felder ableiten.
- Im Feld „phone“ gibt es verschiedene Strukturen für Telefonnummern oder Ausdrücke wie „bitte nicht anrufen“. BDP wird diese Strukturen und deren Häufigkeit analysieren. Der Anwender kann dann entscheiden, welche Stuktur gültig ist und ob Ausdrücke durch ein leeres Feld ersetzt werden sollen.
- In der Spalte „date“ stehen unterschiedlich strukturierte Datumswerte wie „2/1/2010“, 04/2012“ oder „20/04 2014“. BDP erkennt solche Datumswerte, erlaubt die Auswahl verschiedener Date-Format und gleicht Sie an. So könnte das neue Format „dd-mm-yyyy“ heißen und fehlende Angaben durch eine 1 ersetzt werden, also bezogen auf die Beispiele: „02-01-2010“, „01-04-2012“ oder „20-04-2014“.
- Das Ersetzten von Groß/Kleinschreibung ist ebenso vorgesehen, in diesem Fall werden alle „company_name“ durch Großbuchstaben ersetzt.
- Die Reihenfolge der Spalte kann geändert werden.
- Leerfelder können geprüft werden. Falls der Anteil der Zeilen mit leeren Werten ein vordefinierten Anteil übersteigt, wird ein Alarm generiert.
- Eine Dublettenprüfung mit Fuzzy Logik ist ebenfalls möglich. So könnten alle Kombinationen von Vor- und Nachname, die mehr als einmal vorkommen, ebenfalls einen Alarm auslösen.

Um das „transform“ regelmäßig auszuführen (Scheduling der Prozesse), müssen Jobs mit „policies“ erstellt werden (siehe Abb. 2). Dort werden die Quellen und Ziele sowie die Scheduling Information abgelegt.

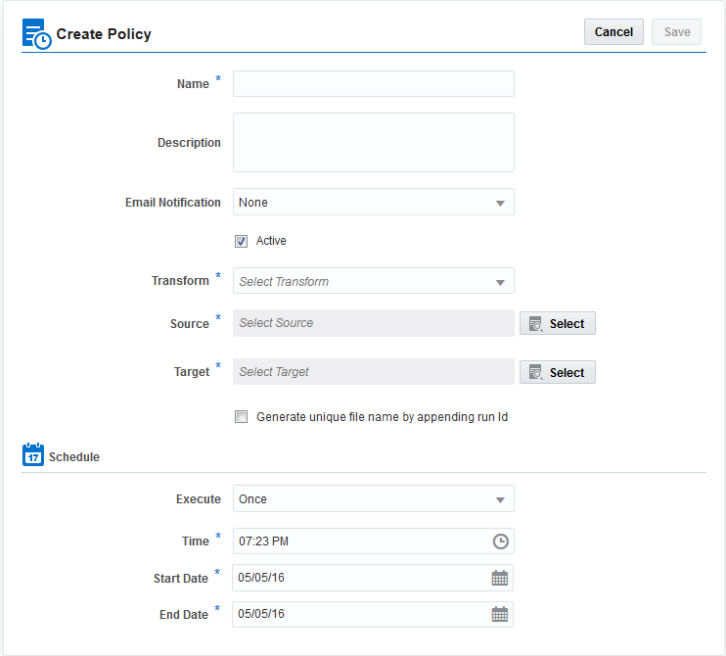


Abb. 6: Policies in BDP

Zusätzlich besteht die Möglichkeit, sich alle Jobs im Überblick anzusehen, um alle Vorgänge zentral administrieren zu können. Die Kommunikation für die Administration zwischen den Anwender und der Cloud erfolgt über E-Mails.

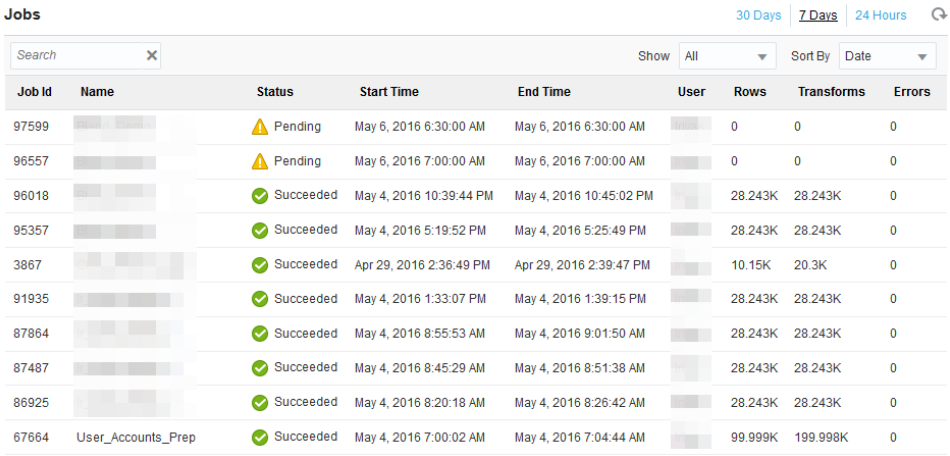


Abb. 7: Jobs in BDP

## 4. Ausblick

BDP ist ein Cloud Applikation, die als Cloud-basierte Plattform as a Service (PaaS) zur Verfügung steht, um Daten in eine in einer Cloud gehostete Umgebung zu laden und zu verarbeiten. Der Service umfasst Funktionen zum Laden, Ergänzen und Anpassen von Daten und validiert Änderungen. Die Benutzeroberfläche ermöglicht, die Arbeiten interaktiv ohne jegliches Coding auszuführen.
Es gibt fortlaufend Erweiterungen, daher ist zu erwarten, dass bis zur DOAG BI Konferenz weitere Feature zur Verfügung stehen.

## 5. Weiterführende Hinweise

http://www.oracle.com/us/products/middleware/data-integration/big-data-prep-cloud-service-ds-2794778.pdf

Übersicht aller BDPCS Tutorials unter: https://apexapps.oracle.com/pls/apex/f?p=44785:1 mit Suche nach „big data preparation“

Using Big Data Preparation Cloud Service Dokumentation: Supported File Types, https://docs.oracle.com/cloud/latest/bdpcs_gs/BDPUG/GUID-DB443D00-3466-4F0E-9584-8F1FB722E0A3.htm#BDPUG-GUID-DB443D00-3466-4F0E-9584-8F1FB722E0A3, Zugriff am 10.05.2016

**Kontaktadresse:**

| | | |
|---|---|---|
| Dr. Holger Dresing | Telefon: | +49 (0) 95787-118 |
| Oracle Deutschland B.V. & Co. KG | Fax: | +49 (0) 95787-118 |
| Thurnithistraße 2 | E-Mail | holger.dresing@oracle.com |
| D-30519 Hannover | Internet: | www.oracle.de |