Data Lake meets KI

Oracle Showcase Projekt

Teil 1:

Erstellung eines Big Data Clusters & Einführung in die Big Data Verarbeitung

daniela.nicola@mt-ag.com, 14. Mai 2018

Einleitung

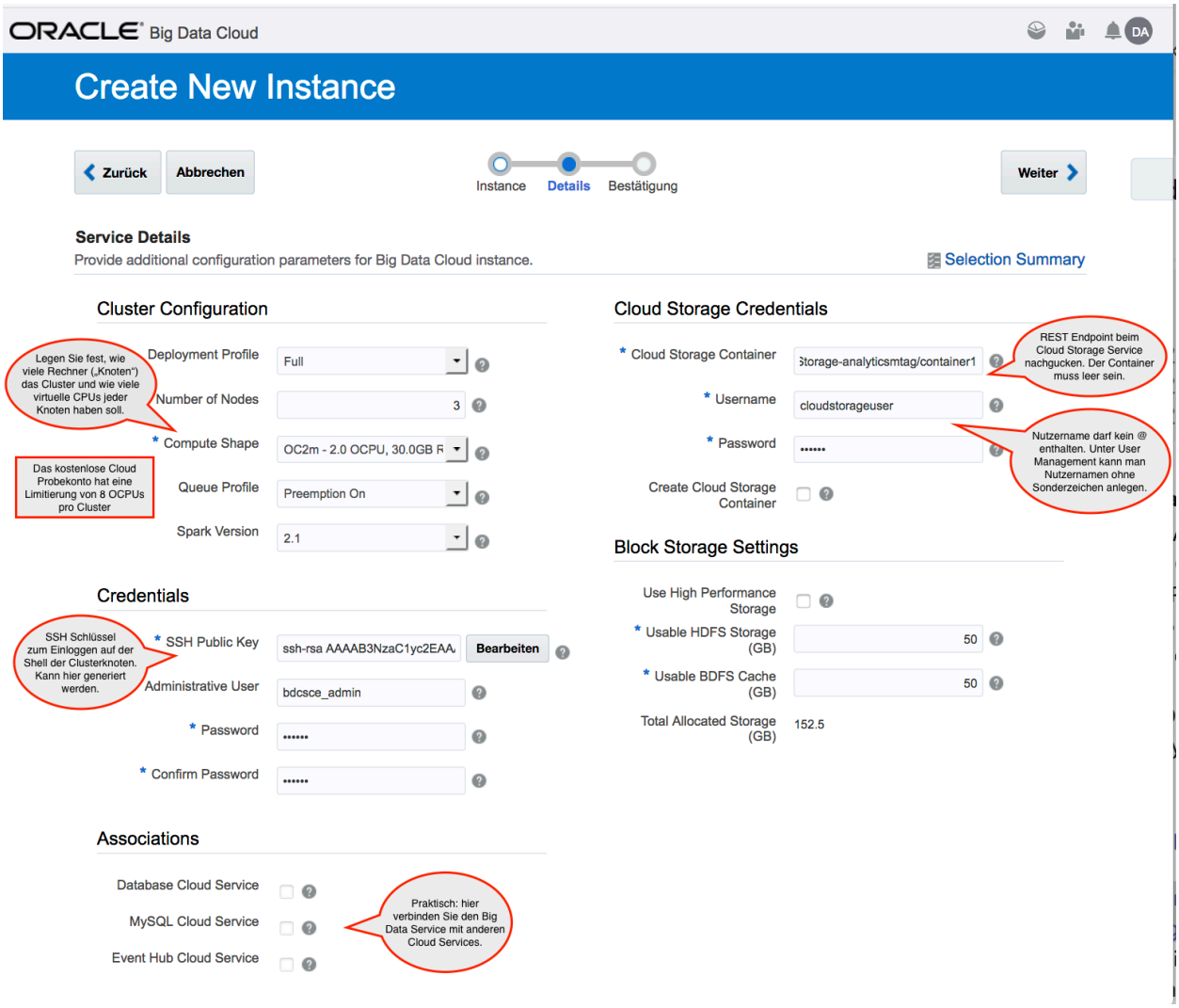
Das Projekt „Data Lake meets KI“ zeigt eine Big Data & KI Musterlösung in der Oracle Cloud. In Teil 1 der Projektbeschreibung zeigen wir Ihnen, wie Sie den Big Data Service in der Oracle Cloud erstellen und dort Daten laden und verarbeiten können.

Was ist der Big Data Service?

Der Big Data Cloud Service ist ein von Oracle gemanagtes Cluster, vorkonfiguriert mit den wichtigsten Komponenten des Apache Hadoop Ökosystems, z.B. Spark, HDFS, Yarn, Tez, Hive, Pig, Ambari, Zeppelin Notebooks. Weitere Komponenten können bei Bedarf hinzugefügt werden.

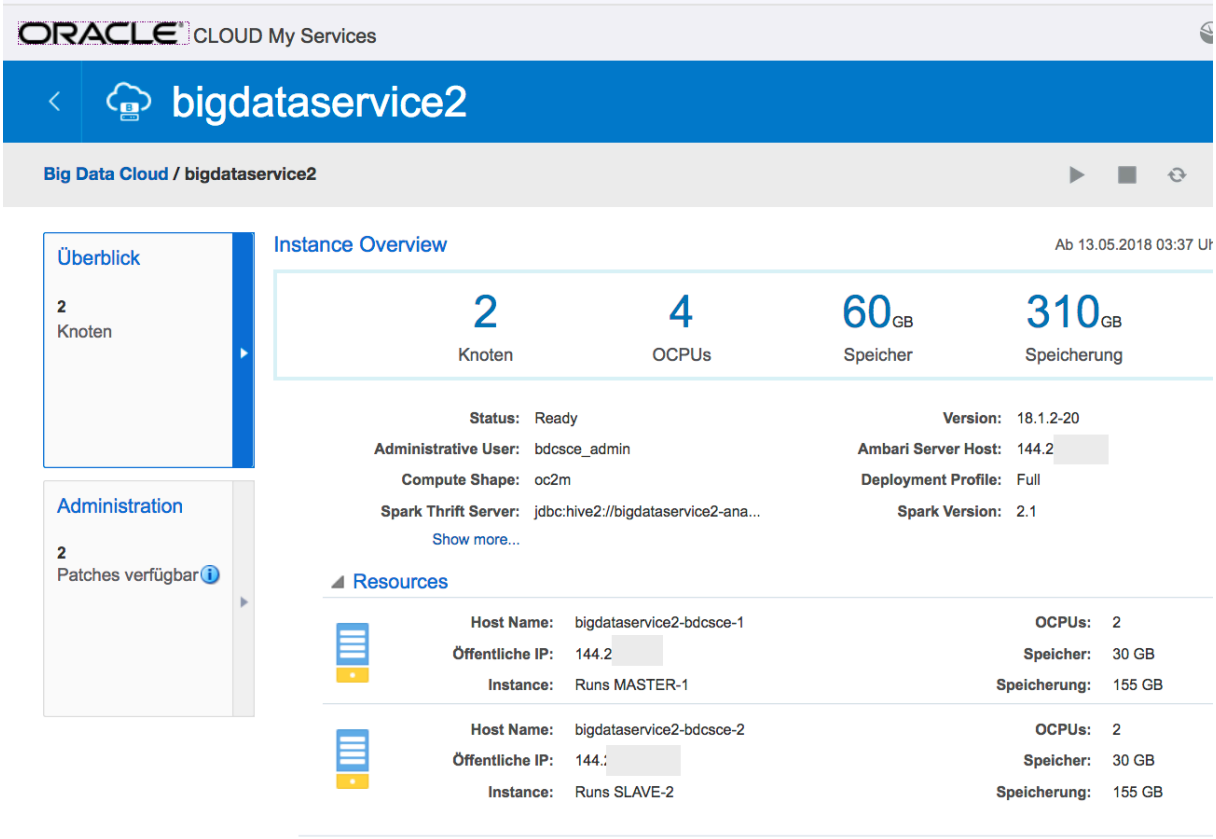
Anlegen des Big Data Service

1. Loggen Sie sich bei der Oracle Cloud ein. Nutzen Sie Ihr existierendes Konto oder erstellen Sie ein kostenloses Probekonto unter https://cloud.oracle.com/de_DE/tryit
2. Erstellen Sie eine Subscription für die Cloud Storage Container Classic. Optional können Sie einen Testcontainer angelegen.
3. Erstellen Sie eine Instanz des Oracle Big Data Services. Hier einige Tipps für die Konfigurationsparameter:



Das war's! In ca. 30 Minuten haben Sie ein Hadoop/Spark Cluster erstellt. So schnell und leicht kann man ein Big Data System im eigenen Rechenzentrum bei weitem nicht erstellen.

Hier ein Beispielcluster mit 2 Knoten:



Dokumentation
https://docs.oracle.com/en/cloud/paas/big-data-compute-cloud/csspc/creating-cluster.html

Was kostet ein Cluster?
Eine Schätzung der Kosten bietet Oracle unter https://cloud.oracle.com/cost-estimator. Um Kosten zu sparen, kann das Cluster bei Nichtbenutzung heruntergefahren werden, oder bei geringem Nutzungsgrad herunterskaliert werden (z.B. weniger Knoten).

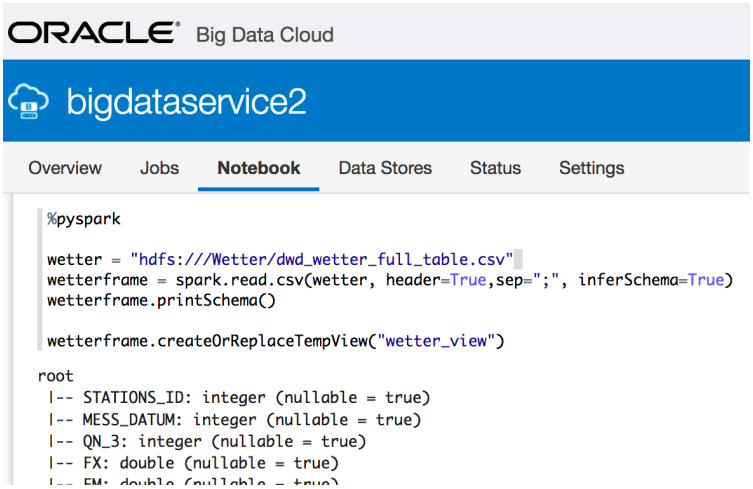
Datentransfer
Kleinere Dateien können über die Big Data Service Webconsole hochgeladen werden. Für größere Dateien bietet Oracle einen speziellen File Transfer Manager an. Für das Laden von Daten in Echtzeit kann der Event Hub Service verwendet werden, der auf Apache Kafka basiert.

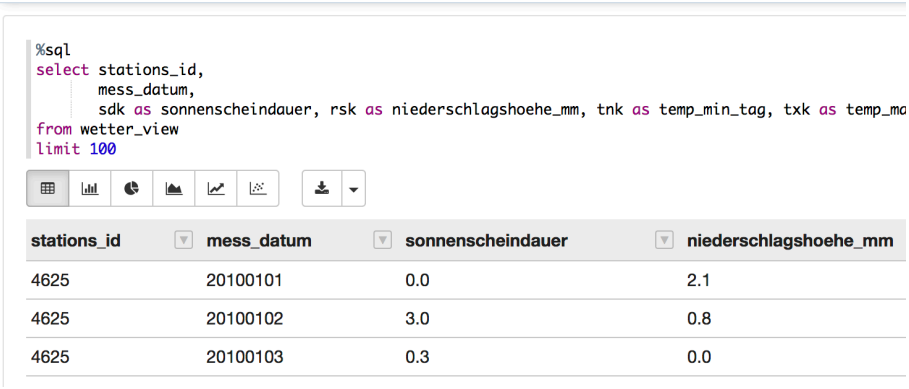
Datenspeicherung
Sie haben die Möglichkeit, Daten lokal auf dem verteilten Dateisystem des Clusters (HDFS) zu speichern. Das hat den Vorteil, dass die Daten direkt auf den Knoten liegen, auf denen sie später per Spark oder MapReduce verarbeitet werden („data locality“). Alternativ kann man Daten im Oracle Cloud Storage ablegen. Letzteres hat den Vorteil, dass die Daten verfügbar sind, auch wenn das Cluster heruntergefahren oder gelöscht wird. Außerdem ist es günstiger als HDFS. Die Performance von Cloud Storage im Vergleich zu HDFS muss von Fall zu Fall untersucht werden. Schnelle Zugriffsmethoden wie das Oracle Big Data Filesystem (BDFS) sollten dabei berücksichtigt werden.

Datenverarbeitung

Für explorative Transformationen von großen Datenmengen empfehlen wir *Zeppelin Notebooks*. Notebooks sind praktische, browserbasierte Editoren, mit denen man Python, Spark, Scala, SQL und Shell Kommandos ausführen, speichern und dokumentieren kann. Notebooks sind Bestandteil des Big Data Services und benutzen Spark als darunterliegende, hoch-performante Ausführungsumgebung.

In der folgenden Abbildung sieht man, wie man in einem Notebook per Spark SQL Anfragen auf einer Textdatei in HDFS ausführen kann:





stations_id	mess_datum	sonnenscheindauer	niederschlagshoehe_mm
4625	20100101	0.0	2.1
4625	20100102	3.0	0.8
4625	20100103	0.3	0.0

Nächste Schritte

In den folgenden Teilen der Projekt Beschreibung zeigen wir Ihnen:

- ETL in der Cloud: Cleansing und Transformation der Daten im Detail
- KI in der Cloud: Machine Learning Algorithmen auf verteilten Cloud Umgebungen
- BI in der Cloud: Visualisierung & Reporting

Kontakt

Kontaktieren Sie die uns, wenn Sie Fragen haben oder Unterstützung bei einem Projekt im Umfeld von Big Data oder Cloud brauchen.

MT AG

Balcke-Dürr-Allee 9
40882 Ratingen
+49 2102 30961-0
info@mt-ag.com