

Working with Apache Hive Design

- Grundsätze
 - LOCATION
 - COMMENT
 - DBPROPERTIES
 - Beispiele
- HIVE Datenbanken
- HIVE Tabellen
- HIVE Datenbanken Design
- HIVE Administration
- Hive-Frontend: HUE
 - Home
 - Query Editor
 - Data Browsers

Grundsätze

Apache Hive ist eine Abstraktionsschicht, die auf dem MapReduce Framework basiert. Es ermöglicht Aufgaben, wie Aggregation, Analysen und Abfragen von Datenmengen, die in HDFS gespeichert sind.

Mit Hive können MapReduce-Programme erzeugt werden ohne das Programmiermodell zu kennen. Dafür wird anstatt einer Datenflusssprache eine SQL-ähnliche Abfragesprache mit dem Namen HiveQL verwendet. Außerdem können in Hive-Tabellenstrukturen angelegt werden. Daher wird es auch als das Hadoop DWH bezeichnet. Allerdings handelt es sich bei Hive-Tabellen lediglich um eine Metadatenschicht über Daten, die sich im HDFS befinden. Dadurch ist beispielsweise das Anlegen von Indizes oder ein Update auf Einzelsatzebene nicht möglich. Im Vergleich zu relationalen Datenbanken hat Hive bei einfachen Abfragen aufgrund von MapReduce hohe Latenzzeiten.

LOCATION

Jede HIVE-Datenbank ist physikalisch nur ein Verzeichnis dessen Namen mit dem Datenbanknamen übereinstimmt. Mit dem Parameter `hiveconf: hive.metastore.warehouse.dir` kann das Standardverzeichnis für die Datenbanken festgelegt bzw. geändert werden. Möchte man eine Datenbank außerhalb des Standardverzeichnisses anlegen, kann man diese mit der `LOCATION` CLAUSE anlegen.

Dies wird verwendet, um für die Units ein eigenes Warehouse Verzeichnis anzulegen (siehe auch HDFS Schema Design).

ACHTUNG: Der Datenbankname muss in der Location mit angegeben werden:

- `LOCATION '/data/dip/global/warehouse/dip_global_stage_tealium.db'`

BEMERKUNG: Die Location wird dann von Hive übernommen und die ACL Rechte werden auf `Hive:Hive` gesetzt.

COMMENT

Zu den Datenbanken können mit der `COMMENT` CLAUSE eine Beschreibung der Datenbank eingefügt werden. Dies kann bei mehreren Teams nützlich sein.

DBPROPERTIES

Analog den Kommentaren, können zusätzliche Eigenschaften der Datenbank als Key-Value Paare hinterlegt werden.

Beispiele

```
DESCRIBE DATABASE dip_global_stage;
-
ALTER DATABASE dip_global_stage SET OWNER USER HIVE;
-
SHOW DATABASES LIKE 'db*';
-
CREATE DATABASE IF NOT EXISTS dip_${Unit}_stage
COMMENT "Stage Area für die Unit ${Unit}"
LOCATION 's3a://dip-${Unit}-test-s3-data-01/warehouse/dip_${Unit}_stage.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_cleanse
COMMENT "Cleanse Area für die Unit <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_cleanse.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_stage
COMMENT "Stage Area für die Unit <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_stage.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_core
COMMENT "Core Area für die Unit <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_core.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_core
COMMENT "Core Area für die Unit <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_core.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_mart_01
COMMENT "Core Area für die Unit <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_mart_01.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_work
COMMENT "Work Area für die Unit <UNIT>"
LOCATION '/data/dip/welt/warehouse/dip_<UNIT>_work.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_default
COMMENT "Default database do not delete"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_default.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
```

```

'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_mine
COMMENT "Mining database for <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_mine.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);
CREATE DATABASE IF NOT EXISTS dip_<UNIT>_export
COMMENT "Export database for <UNIT>"
LOCATION '/data/dip/<UNIT>/warehouse/dip_<UNIT>_export.db'
WITH DBPROPERTIES(
'created_date' = '2017-03-07',
'created_by' = 'Stephan Semmler',
'Email' = 'stephan.semmler@axelspringer.de'
);

```

HIVE Datenbanken

Alle HIVE Datenbanken greifen auf einen Metadastore zu. Deshalb müssen die Datenbank-Namen eindeutig sein, auch wenn sie in verschiedenen Verzeichnissen abgelegt sind. Tabellen hingegen müssen nur innerhalb einer Datenbank eindeutig sein. Deshalb wird jeder Datenbank ein Unit-Präfix, Prozess-Präfix sowie eine Source-Präfix hinzugefügt „dip_<Unit><Prozess><Quelle>“, z.B.:

- dip_global_stage_eventstore

HIVE Tabellen

Wenn eine externe Tabelle erstellt wird, müssen für den Benutzer die Hive Zugriffs-Rechte auch auf dieses Verzeichnis gesetzt werden.

HIVE Datenbanken Design

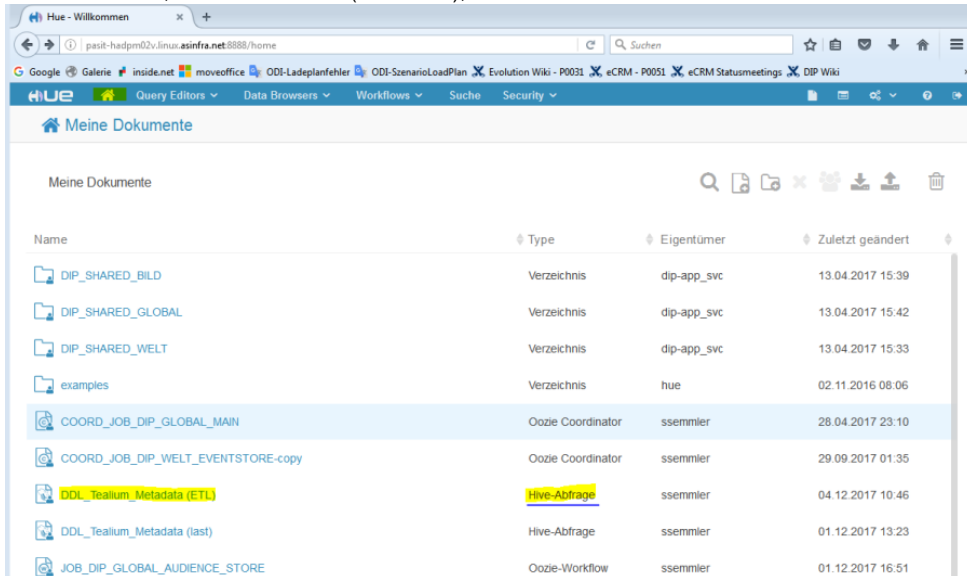
HIVE Administration

Die HIVE Datenbanken werden durch das CDP-Admin-Team angelegt und verwaltet. Innerhalb der Datenbanken können, je nach interner Rechtevergabe der Units, Tabellen durch die User angelegt bearbeitet oder gelöscht werden. Ausnahme sind Tabellen, die in Absprache mit den Units vom CDP-Admin-Team bewirtschaftet werden.

Hive-Frontend: HUE

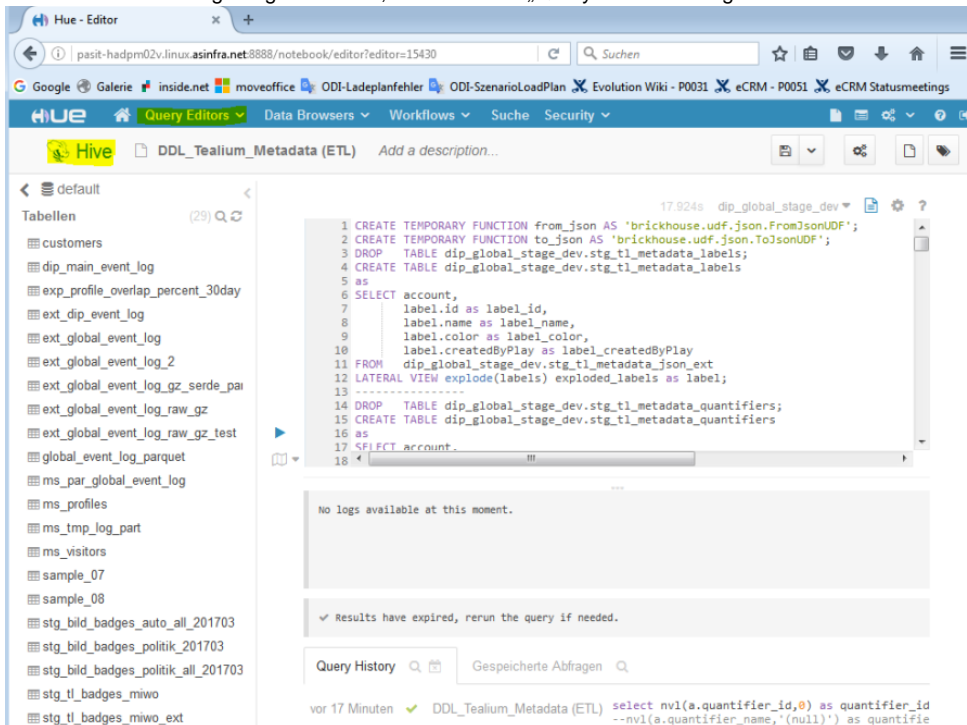
Home

Im Home-Bereich sind die Dokumente hinterlegt, die bereits erzeugt und gespeichert wurden. Dabei handelt es sich u.a. um Hive-Abfragen, Editor-Workflows, Editor-Coordinator(Scheduler),



Name	Type	Eigentümer	Zuletzt geändert
DIP_SHARED_BILD	Verzeichnis	dip-app_svc	13.04.2017 15:39
DIP_SHARED_GLOBAL	Verzeichnis	dip-app_svc	13.04.2017 15:42
DIP_SHARED_WELT	Verzeichnis	dip-app_svc	13.04.2017 15:33
examples	Verzeichnis	hue	02.11.2016 08:06
COORD_JOB_DIP_GLOBAL_MAIN	Oozie Coordinator	ssemmler	28.04.2017 23:10
COORD_JOB_DIP_WELT_EVENTSTORE-copy	Oozie Coordinator	ssemmler	29.09.2017 01:35
DDL_Tealium_Metadata (ETL)	Hive-Abfrage	ssemmler	04.12.2017 10:46
DDL_Tealium_Metadata (last)	Hive-Abfrage	ssemmler	01.12.2017 13:23
JOB_DIP_GLOBAL_AUDIENCE_STORE	Oozie-Workflow	ssemmler	01.12.2017 16:51

Wenn eine Hive-Abfrage angeklickt wird, dann wird der „Query Editor“ Hive geöffnet.



Hue - Editor

pasit-hadpm02v.linux.asinfra.net:8888/notebook/editor?editor=15430

Hive DDL_Tealium_Metadata (ETL) Add a description...

Tabellen (29) 🔍

- customers
- dip_main_event_log
- exp_profile_overlap_percent_30day
- ext_dip_event_log
- ext_global_event_log
- ext_global_event_log_2
- ext_global_event_log_gz_serde_pai
- ext_global_event_log_raw_gz
- ext_global_event_log_raw_gz_test
- global_event_log_parquet
- ms_par_global_event_log
- ms_profiles
- ms_tmp_log_part
- ms_visitors
- sample_07
- sample_08
- stg_bild_badges_auto_all_201703
- stg_bild_badges_politik_201703
- stg_bild_badges_politik_all_201703
- stg_ti_badges_miwo
- stg_ti_badges_miwo_ext

```
1 CREATE TEMPORARY FUNCTION from_json AS 'brickhouse.udf.json.FromJsonUDF';
2 CREATE TEMPORARY FUNCTION to_json AS 'brickhouse.udf.json.ToJsonUDF';
3 DROP TABLE dip_global_stage_dev.stg_ti_metadata_labels;
4 CREATE TABLE dip_global_stage_dev.stg_ti_metadata_labels
5 AS
6 SELECT account,
7        label.id as label_id,
8        label.name as label_name,
9        label.color as label_color,
10       label.createdByPlay as label_createdByPlay
11 FROM   dip_global_stage_dev.stg_ti_metadata_json_ext
12 LATERAL VIEW explode(labels) exploded_labels as label;
13 -----
14 DROP TABLE dip_global_stage_dev.stg_ti_metadata_quantifiers;
15 CREATE TABLE dip_global_stage_dev.stg_ti_metadata_quantifiers
16 AS
17 SELECT account,
```

No logs available at this moment.

✓ Results have expired, rerun the query if needed.

Query History 🔍 Gespeicherte Abfragen 🔍

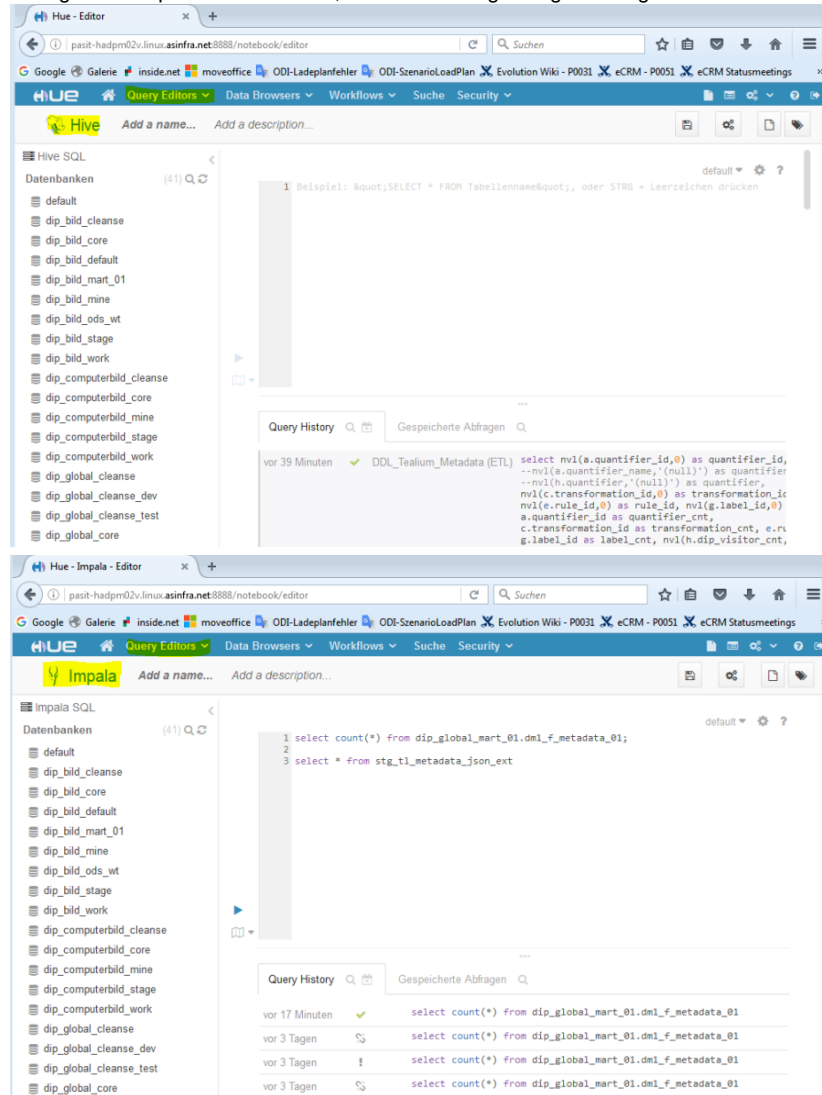
vor 17 Minuten ✓ DDL_Tealium_Metadata (ETL) select nvl(s.quantifier_id,0) as quantifier_id
--nvl(s.quantifier_name,'(null)') as quantifie

Query Editor

Unter „Query Editor“ sind folgende Editoren gelistet:

- Hive
- Impala
- DB-Abfrage
- Pig (wird nicht genutzt)
- Job Designer

Abfragen mit Impala sind schneller, als mit Hive. Pig wird gar nicht genutzt.



Beim öffnen der Editoren wird standardmäßig die default Datenbank geöffnet.

Data Browsers

Unter „Data Browsers“ befinden sich:

- Metastore Manager
- hBase

The top screenshot shows the Hue Metastore Manager interface. The sidebar on the left lists various tables, including 'customers', 'dip_main_event_log', 'evg_profile_overlap_percent_30day', 'ext_dip_event_log', 'ext_global_event_log', 'ext_global_event_log_2', 'ext_global_event_log_gr_serde_par', 'ext_global_event_log_raw_gr', 'global_event_log_parquet', 'ms_par_global_event_log', 'ms_profiles', 'ms_tmp_log_part', 'ms_visitors', 'sample_07', 'sample_08', 'stg_bild_badges_auto_all_201703', 'stg_bild_badges_politik_201703', 'stg_bild_badges_politik_all_201703', 'stg_til_badges_miwo', and 'stg_til_badges_miwo_ext'. The main view displays the 'dip_bild_cleanse' database, showing its statistics and a list of tables. The tables listed include 'cls_visitor_h', 'cls_visitor_s_audience_audiences', 'cls_visitor_s_audience_badges', 'cls_visitor_s_audience_dates', 'cls_visitor_s_audience_flags', 'cls_visitor_s_audience_metrics_sets', 'cls_visitor_s_audience_metrics', 'cls_visitor_s_audience_properties', and 'cls_visitor_s_audience_property_sets'.

The bottom screenshot shows the Hue HBase-Browser interface. The sidebar on the left lists various tables, including 'employee' and 'hbase_table'. The main view displays the 'employee' table, showing its statistics and a list of tables. The tables listed include 'employee' and 'hbase_table'.