Workflow EventDB

Grundsätze

Die Daten die für die AudienceDB werden durch Tealium generiert und im Audiencestore persitiert.

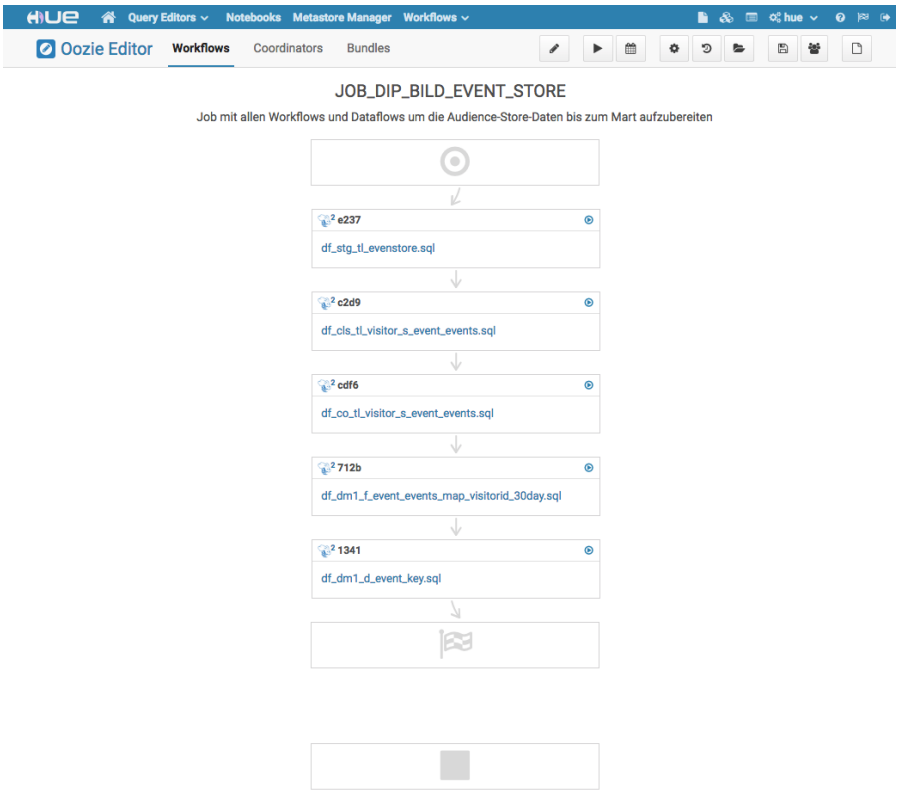
Quellen

Jede Unit hat eine eigene AudienceDB und die Daten werden in einem unitspezifischen nicht partitioniertten Audiencestore gespeichert. (siehe auch unter Schnittstellen)
Die Daten werden aus dem Audiencestore mit dem AWS-CLI-Tool abgeholt. Das entsprechende Shell-Script wird über die Crontab auf dem Managment-Server aufgerufen und im eigen Data Bucket von CDP partitioniert gespeichert.

```
s3://dip-global-test-s3-app-01/aws/bin/dip-global-test-emr-imp-01.sh
s3://dip-global-test-s3-app-01/emr/bin/step/dip-global-test-emr-step-copy-tealium-audiences.sh
s3://dip-global-test-s3-app-01/emr/bin/step/dip-global-test-emr-step-copy-tealium-events.sh
```

Workflow mit Oozie

Anschließend werden die Daten mit Hive aufbereitet. Dazu wurde ein Workflow mit Oozie erstellt, der alle einzelnen Dataflows orchestrisiert.



Dataflows

Die einzelnen Dataflows bestehen aus HiveSQL.

Stage

Die Rohdaten liegen im Json-Format vor und werden über einen external Table aus S3 in den HDFS-Bereich geladen, da dies eine schnellere Bearbeitung ermöglicht.

Cleanse

Die Json-Rohdaten werden in Tabellen-Struktur mit komplexen Datatypes überführt.

Core

Die Tagesdaten in Cleanse werden in die historisierte und partitionierte Struktur kopiert.

Data Mart

Die Core Daten werden für spezielle Sichten aus den Fachbereichen aufbereitet. (z.B. AudienceDB)