Solution – Exercise IV

HiveQL, HDFS/Hive Partitioning, HiveServer2

Solution

Prerequisites:

- Start Gcloud instance
- Pull and start Docker image (`marcelmittelstaedt/hiveserver_base:latest`)
- Start Hadoop Cluster
- Start HiveServer2
- Download, Install and Configure JDBC Rich-client:
    - e.g. DBeaver,
    - SquirrelSQL,
    - …
- Execute all preparation and example tasks of previous HandsOn slides of last lecture

Solution

Exercise IV:

2.1 Create table **`name_basics_partitioned`** partitioned by column `partition_is_alive`:

```
CREATE EXTERNAL TABLE IF NOT EXISTS name_basics_partitioned (
          nconst STRING,
          primary_name STRING,
          birth_year INT,
          death_year STRING,
          primary_profession STRING,
          known_for_titles STRING
) PARTITIONED BY (partition_is_alive STRING)
STORED AS ORCFILE LOCATION '/user/hadoop/imdb/actors_partitioned';
```

Solution

Exercise IV:

2.2 Use static partitioning to create and fill partition ‘alive‘

```
INSERT OVERWRITE TABLE name_basics_partitioned
partition(partition_is_alive='alive')
SELECT
          a.nconst,
          a.primary_name,
          a.birth_year,
          a.death_year,
          a.primary_profession,
          a.known_for_titles
FROM name_basics a WHERE a.death_year IS NULL
```

Solution

Exercise IV:

2.3 Use static partitioning to create and fill partition ‘dead‘

```
INSERT OVERWRITE TABLE name_basics_partitioned
partition(partition_is_alive='dead')
SELECT
          a.nconst,
          a.primary_name,
          a.birth_year,
          a.death_year,
          a.primary_profession,
          a.known_for_titles
FROM name_basics a WHERE a.death_year IS NOT NULL
```

Solution

Exercise IV:

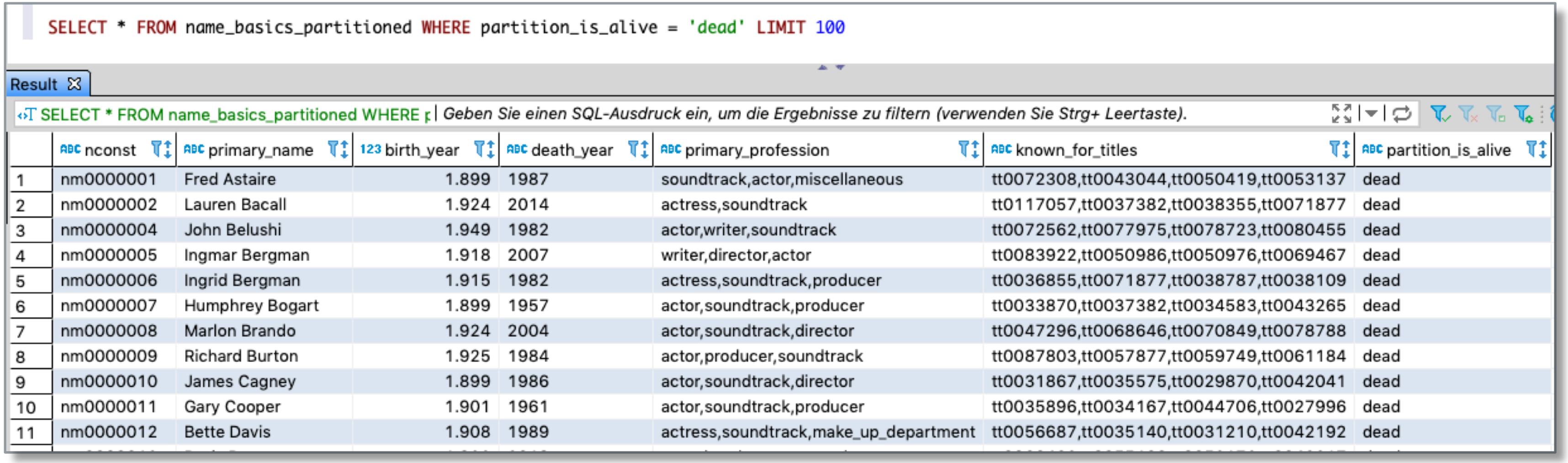
2.4 Check Results:

```
hadoop fs -ls /user/hadoop/imdb/actors_partitioned
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:18 /user/hadoop/imdb/actors_partitioned/partition_is_alive=alive
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:18 /user/hadoop/imdb/actors_partitioned/partition_is_alive=dead
```

Solution

Exercise IV:

2.4 Check Results:



```
SELECT * FROM name_basics_partitioned WHERE partition_is_alive = 'dead' LIMIT 100
```

Result

SELECT * FROM name_basics_partitioned WHERE p | Geben Sie einen SQL-Ausdruck ein, um die Ergebnisse zu filtern (verwenden Sie Strg+ Leertaste).

	nconst	primary_name	birth_year	death_year	primary_profession	known_for_titles	partition_is_alive
1	nm0000001	Fred Astaire	1.899	1987	soundtrack,actor,miscellaneous	tt0072308,tt0043044,tt0050419,tt0053137	dead
2	nm0000002	Lauren Bacall	1.924	2014	actress,soundtrack	tt0117057,tt0037382,tt0038355,tt0071877	dead
3	nm0000004	John Belushi	1.949	1982	actor,writer,soundtrack	tt0072562,tt0077975,tt0078723,tt0080455	dead
4	nm0000005	Ingmar Bergman	1.918	2007	writer,director,actor	tt0083922,tt0050986,tt0050976,tt0069467	dead
5	nm0000006	Ingrid Bergman	1.915	1982	actress,soundtrack,producer	tt0036855,tt0071877,tt0038787,tt0038109	dead
6	nm0000007	Humphrey Bogart	1.899	1957	actor,soundtrack,producer	tt0033870,tt0037382,tt0034583,tt0043265	dead
7	nm0000008	Marlon Brando	1.924	2004	actor,soundtrack,director	tt0047296,tt0068646,tt0070849,tt0078788	dead
8	nm0000009	Richard Burton	1.925	1984	actor,producer,soundtrack	tt0087803,tt0057877,tt0059749,tt0061184	dead
9	nm0000010	James Cagney	1.899	1986	actor,soundtrack,director	tt0031867,tt0035575,tt0029870,tt0042041	dead
10	nm0000011	Gary Cooper	1.901	1961	actor,soundtrack,producer	tt0035896,tt0034167,tt0044706,tt0027996	dead
11	nm0000012	Bette Davis	1.908	1989	actress,soundtrack,make_up_department	tt0056687,tt0035140,tt0031210,tt0042192	dead

Solution

Exercise IV:

3.1 Create table **imdb_movies_and_ratings_partitioned** partitioned by column partition_year using fields of table title_basics and title_ratings:

```
CREATE TABLE IF NOT EXISTS imdb_movies_and_ratings_partitioned (
           tconst STRING,
           title_type STRING,
           primary_title STRING,
           original_title STRING,
           is_adult DECIMAL(1,0),
           start_year DECIMAL(4,0),
           end_year STRING,
           runtime_minutes INT,
           genres STRING,
           average_rating DECIMAL(2,1),
           num_votes BIGINT
) PARTITIONED BY (partition_year int) STORED AS ORCFILE LOCATION '/user/hadoop/imdb/
movies_and_ratings_partitioned';
```

Solution

Exercise IV:

3.2 Use dynamic partitioning to create and fill partition `partition_year`:

```
SET hive.exec.dynamic.partition.mode=nonstrict;
INSERT OVERWRITE TABLE imdb_movies_and_ratings_partitioned partition(partition_year)
SELECT
            tb.tconst,
            tb.title_type,
            tb.primary_title,
            tb.original_title,
            tb.is_adult,
            tb.start_year,
            tb.end_year,
            tb.runtime_minutes,
            tb.genres,
            tr.average_rating,
            tr.num_votes,
            tb.start_year
FROM title_basics tb JOIN title_ratings tr ON (tb.tconst = tr.tconst)
```

Solution

Exercise IV:

3.3 Check Results:

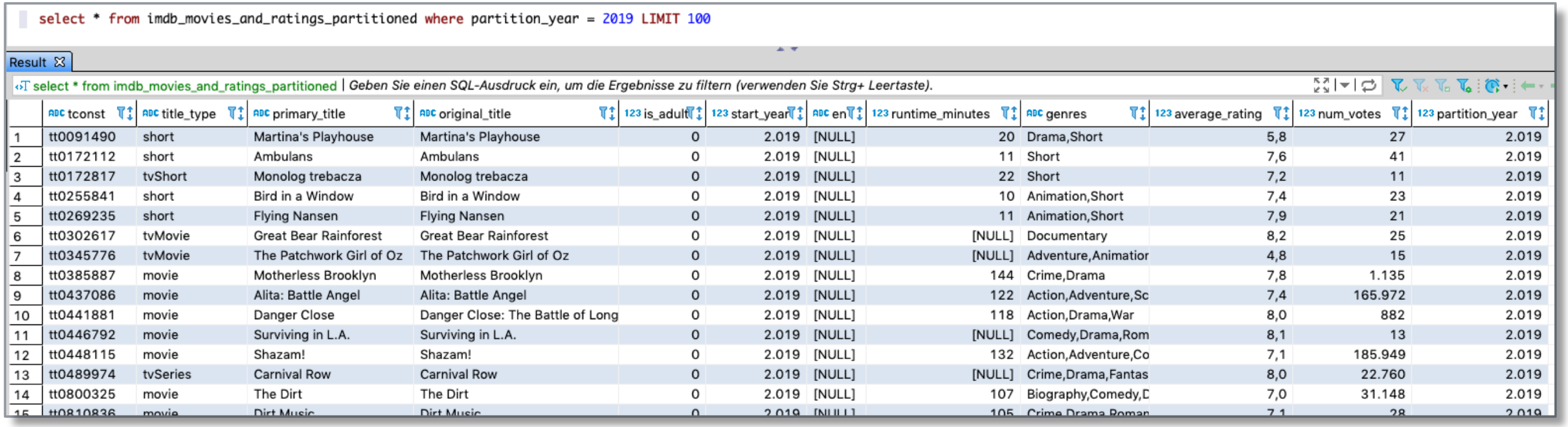
```
hadoop fs -ls /user/hadoop/imdb/movies_and_ratings_partitioned

[…]
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1874
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1878
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1881
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1883
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1885
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1887
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1888
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1889
drwxr-xr-x   - hadoop supergroup          0 2019-10-20 18:27 /user/hadoop/imdb/movies_and_ratings_partitioned/partition_year=1890
[…]
```

Solution

Exercise IV:

3.3 Check Results:



```
select * from imdb_movies_and_ratings_partitioned where partition_year = 2019 LIMIT 100
```

Result

select * from imdb_movies_and_ratings_partitioned | *Geben Sie einen SQL-Ausdruck ein, um die Ergebnisse zu filtern (verwenden Sie Strg+ Leertaste).*

	tconst	title_type	primary_title	original_title	is_adult	start_year	en	runtime_minutes	genres	average_rating	num_votes	partition_year
1	tt0091490	short	Martina's Playhouse	Martina's Playhouse	0	2.019	[NULL]	20	Drama,Short	5,8	27	2.019
2	tt0172112	short	Ambulans	Ambulans	0	2.019	[NULL]	11	Short	7,6	41	2.019
3	tt0172817	tvShort	Monolog trebacza	Monolog trebacza	0	2.019	[NULL]	22	Short	7,2	11	2.019
4	tt0255841	short	Bird in a Window	Bird in a Window	0	2.019	[NULL]	10	Animation,Short	7,4	23	2.019
5	tt0269235	short	Flying Nansen	Flying Nansen	0	2.019	[NULL]	11	Animation,Short	7,9	21	2.019
6	tt0302617	tvMovie	Great Bear Rainforest	Great Bear Rainforest	0	2.019	[NULL]	[NULL]	Documentary	8,2	25	2.019
7	tt0345776	tvMovie	The Patchwork Girl of Oz	The Patchwork Girl of Oz	0	2.019	[NULL]	[NULL]	Adventure,Animatior	4,8	15	2.019
8	tt0385887	movie	Motherless Brooklyn	Motherless Brooklyn	0	2.019	[NULL]	144	Crime,Drama	7,8	1.135	2.019
9	tt0437086	movie	Alita: Battle Angel	Alita: Battle Angel	0	2.019	[NULL]	122	Action,Adventure,Sc	7,4	165.972	2.019
10	tt0441881	movie	Danger Close	Danger Close: The Battle of Long	0	2.019	[NULL]	118	Action,Drama,War	8,0	882	2.019
11	tt0446792	movie	Surviving in L.A.	Surviving in L.A.	0	2.019	[NULL]	[NULL]	Comedy,Drama,Rom	8,1	13	2.019
12	tt0448115	movie	Shazam!	Shazam!	0	2.019	[NULL]	132	Action,Adventure,Co	7,1	185.949	2.019
13	tt0489974	tvSeries	Carnival Row	Carnival Row	0	2.019	[NULL]	[NULL]	Crime,Drama,Fantas	8,0	22.760	2.019
14	tt0800325	movie	The Dirt	The Dirt	0	2.019	[NULL]	107	Biography,Comedy,D	7,0	31.148	2.019
15	tt0810836	movie	Dirt Music	Dirt Music	0	2.019	[NULL]	105	Crime,Drama,Roman	7,1	28	2.019