Project documentation

Business Intelligence Project Lab 3

docker run --hostname=quickstart.cloudera --privileged=true -t -i -p 8888 cloudera/quickstart:latest /usr/bin/docker-quickstart

Run `gradle runAllMapReduceJobs` in the project folder

map reduce 1

Implementation:

- gradle as build system (gradle)
- mapper to parse the file and map to (key: product, value: 1)
- reducer which produces (key: product, value: avg(product)) which calculates the arithmetic mean per product

Scalability:

- expected speedup when run on multiple machines
  - a bit less than linear due to cooridnation overhead

Result: first 15 scores per category

Amazon Instant Video:

```
B000GFDAUG	5.0
B000GIOPK2	4.351351351351352
B000GIPKWY	3.857142857142857
B000GJUQ7M	3.6666666666666665
B000GK0NBK	3.5
B000GK51HG	5.0
B000GK6NFK	4.389830508474576
B000GK7DPY	3.6666666666666665
B000GOTJGG	2.3333333333333335
B000GOV10S	5.0
B000GOW7RE	5.0
B000GOW9B8	3.0
B000GOYLNC	5.0
B000GP0T82	3.0
B000GP38JE	5.0
```

Sports and Outdoors:

```
0000031852	3.857142857142857
0000031895	3.6666666666666665
0000031909	2.6666666666666665
0000032034	3.6666666666666665
0000032050	4.75
0000032069	4.0
0188477284	2.0
0531904822	5.0
```

```
059445039X    3.0
060791548X    3.0
0607968699    3.0
0615302939    5.0
0615329020    4.666666666666667
0615375790    5.0
0615532209    5.0
```

Video Games:

```
0078764343    4.666666666666667
043933702X    4.0
0439339960    3.0
0439339987    5.0
0439342260    4.0
0439374391    5.0
0439394422    4.0
043940133X    3.2857142857142856
0439573947    4.666666666666667
0439591295    5.0
0439591368    3.75
0439591538    2.5
0439671418    5.0
0439715571    4.0
0439773660    2.0
```

map reduce 2 sentiment analysis

Implementation:

- gradle as build system
- mapper to parse the file and map to (key: product, value1: numberPosWords, value2: numberNegWords)
- reducer which produces (key: product, value: sentiment(product, #posWords, #negWords)) which calculates the sentiment score per product

Questions:

- How many invocations of map reduce are there:
  - Amazon Instant Video: 37126 and 1685
  - Sports and Outdoors: 296337 and 18357
  - Video Games: 231780 and 10672
- where is most of the runtime spent
  - parsing the json, disk IO
- what is the expected speedup when run on multiple machines
  - less than linear due to shuffle overhead (network is slow)

Result: first 10 products per category

Sports and Outdoors:

```
1881509818    SentimentWritable [positive=17, negative=10, sentiment=0.25925925925925924]
2094869245    SentimentWritable [positive=39, negative=8, sentiment=0.6595744680851063]
7245456259    SentimentWritable [positive=28, negative=7, sentiment=0.6]
7245456313    SentimentWritable [positive=497, negative=168, sentiment=0.49473684210526314]
B000002NUS    SentimentWritable [positive=53, negative=28, sentiment=0.30864197530864196]
```

```
B00000EL25    SentimentWritable [positive=26, negative=22, sentiment=0.08333333333333333]
B00000IURU    SentimentWritable [positive=32, negative=5, sentiment=0.7297297297297297]
B00000IUX5    SentimentWritable [positive=38, negative=4, sentiment=0.8095238095238095]
B00000J6JO    SentimentWritable [positive=148, negative=73, sentiment=0.3393665158371041]
B0000224UE    SentimentWritable [positive=100, negative=34, sentiment=0.4925373134328358]
```

Amazon Instant Video:

```
B000H00VBQ    SentimentWritable [positive=31, negative=39, sentiment=-0.11428571428571428]
B000H0X79O    SentimentWritable [positive=12, negative=3, sentiment=0.6]
B000H29TXU    SentimentWritable [positive=9, negative=3, sentiment=0.5]
B000H2DMME    SentimentWritable [positive=32, negative=12, sentiment=0.45454545454545453]
B000H4YNM0    SentimentWritable [positive=100, negative=88, sentiment=0.06382978723404255]
B000HAB4NK    SentimentWritable [positive=212, negative=155, sentiment=0.1553133514986376]
B000HKWE3O    SentimentWritable [positive=11, negative=2, sentiment=0.6923076923076923]
B000HZEHL6    SentimentWritable [positive=104, negative=82, sentiment=0.11827956989247312]
B000I5PVD8    SentimentWritable [positive=27, negative=9, sentiment=0.5]
B000I5Q0ZG    SentimentWritable [positive=13, negative=4, sentiment=0.5294117647058824]
```

Video games:

```
0700099867    SentimentWritable [positive=158, negative=110, sentiment=0.1791044776119403]
6050036071    SentimentWritable [positive=33, negative=15, sentiment=0.375]
7100027950    SentimentWritable [positive=50, negative=61, sentiment=-0.0990990990990991]
7293000936    SentimentWritable [positive=18, negative=13, sentiment=0.16129032258064516]
8176503290    SentimentWritable [positive=58, negative=31, sentiment=0.30337078651685395]
907843905X    SentimentWritable [positive=33, negative=20, sentiment=0.24528301886792453]
9625990674    SentimentWritable [positive=76, negative=21, sentiment=0.5670103092783505]
9861019731    SentimentWritable [positive=36, negative=5, sentiment=0.7560975609756098]
9882155456    SentimentWritable [positive=91, negative=52, sentiment=0.2727272727272727]
B000003SQQ    SentimentWritable [positive=50, negative=27, sentiment=0.2987012987012987]
```

hive

RESULTS FOR ALL QUERIES

How many movies are there in total in the dataset?

```
32204001
```

1 Job

Map Reduce = sum

How many movies in the dataset belong to the "Film-Noir" genre?

```
233
```

1 job

Where, Sum

Which are the 10 most frequently assigned tags (by users, i.e., from the tags table)?

```
1    sci-fi    3384
2    based on a book    3281
3    atmospheric    2917
4    comedy    2779
```

```
5    action    2657
6    surreal    2427
7    BD-R    2334
8    twist ending    2323
9    funny    2072
10    dystopia    1991
```

Job 1: Group and calculate SUM Job 2: Sort and Limit

Which 10 movies were the most controversial in 2015 (i.e., had the highest variance in ratings between 2015/01/01 and 2015/12/31)?

```
1    45533    4
2    6051    4
3    2298    4
4    101971    4
5    128425    4
6    126927    4
7    3905    4
8    72360    4
9    128173    4
10    128169    4
```

2 jobs

Job 1: Group, Variance (AVG) Job 2: Variance, Order

Which movies (titles) are the 10 most frequently tagged and how often have they been tagged?

```
1    Pulp Fiction (1994)    1994
2    Fight Club (1999)    1779
3    Inception (2010)    1552
4    "Matrix    1430
5    "Shawshank Redemption    1339
6    Eternal Sunshine of the Spotless Mind (2004)    1240
7    Donnie Darko (2001)    1177
8    Memento (2000)    1168
9    "Silence of the Lambs    1100
10    Avatar (2009)    995
```

2 jobs

Job 1: Group and calculate SUM Job 2: Join, Sort and Limit

Which 15 movies (titles) have been most frequently tagged with the label "mars"?

```
1    Mars Attacks! (1996)    34
2    "War of the Worlds    25
3    Total Recall (2012)    10
4    Capricorn One (1978)    9
5    Total Recall (1990)    9
6    Martian Child (2007)    6
7    It Came from Outer Space (1953)    4
8    Mission to Mars (2000)    4
9    "Day the Earth Stood Still    4
10    RocketMan (a.k.a. Rocket Man) (1997)    3
11    "6th Day    3
12    Red Planet (2000)    2
13    Destination Moon (1950)    2
14    Impostor (2002)    1
15    Doom (2005)    1
```

```
15    Doom (2005)    1
```

2 jobs

Job 1: Where, group and calculate SUM Job 2: Join, Sort and Limit

Which are the 10 best-rated movies (on average; list titles) with more than 1000 ratings?

```
1    Pulp Fiction (1994)    4.2625    67310
2    Forrest Gump (1994)    4.1235    66172
3    "Shawshank Redemption    4.539    63366
4    "Silence of the Lambs    4.2695    63299
5    Jurassic Park (1993)    3.7567    59715
6    Star Wars: Episode IV - A New Hope (1977)    4.2798    54502
7    Braveheart (1995)    4.1327    53769
8    Terminator 2: Judgment Day (1991)    4.0214    52244
9    "Matrix    4.3169    51334
10    Schindler's List (1993)    4.4015    50054
```

2 jobs

Job 1: group and calculate SUM Job 2: Having, Join, Sort and Limit

Which are the highest-rated "Film-Noir" movies with more than 10 ratings (average rating; movies with genre "Film-Noir", max. 10)?

```
1    L.A. Confidential (1997)    4.1845    26836
2    Sin City (2005)    4.0856    15481
3    Chinatown (1974)    4.2995    15310
4    Dark City (1998)    3.9311    11759
5    Mulholland Drive (2001)    3.986    9307
6    Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)    4.3537    6525
7    Miller's Crossing (1990)    4.0957    6060
8    Strangers on a Train (1951)    4.2631    5154
9    Blood Simple (1984)    4.1587    5053
10    Notorious (1946)    4.2847    4932
```

2 jobs

Job 1: where, group and calculate SUM Job 2: Having, Join, Sort and Limit

What are the 15 most relevant genome tags for the movie "Toy Story (1995)" (movieId=1)?

```
1    toys    0.99924999999999997
2    computer animation    0.99849999999999994
3    pixar animation    0.996
4    kids and family    0.99075000000000002
5    animation    0.9857499999999999
6    kids    0.97924999999999995
7    pixar    0.96675
8    children    0.96425000000000005
9    cartoon    0.95649999999999991
10    imdb top 250    0.94199999999999995
11    animated    0.93324999999999991
12    childhood    0.92625000000000002
13    great movie    0.92074999999999996
14    disney animated feature    0.91375000000000006
15    friendship    0.91175000000000006
```

1 job

Where, Join, Order, Limit (kein Group)

Which are the 10 most relevant movies for Vienna (i.e., with the highest genome tag relevance rating for the tag "vienna")?

```
1    "Third Man    0.98750000000000004
2    Johnny Guitar (1954)    0.96649999999999991
3    Before Sunrise (1995)    0.96124999999999994
4    Before Sunset (2004)    0.95700000000000007
5    Before Midnight (2013)    0.91100000000000003
6    "Night Porter    0.89349999999999996
7    "Illusionist    0.84525000000000006
8    Amadeus (1984)    0.84149999999999991
9    "Foreign Affair    0.73224999999999996
10    Love in the Afternoon (1957)    0.66900000000000004
```

1 job

Where, Join, Order, Limit (kein Group)

DESCRIPTION OF YOUR UNDERSTANDING OF WHAT HAPPENS BEHIND THE SCENES

including a discuss on how many MR jobs your queries are translated into and why. Finally, also comment on what scale-up you would expect when running your queries on a real cluster in paralle

Scalability:

- expected speedup when run on multiple machines
  - a bit less than linear due to cooridnation overhead