Business Intelligence 2017W

Assignment 2

Group 53:

Maximilian Moser (01326252)
Roman Tonigold (01327192)
Wolfgang Weintritt (01327191)

1. MapReduce

We chose the following Amazon review datasets:

- Electronics
- Video Games
- Baby

Documentation of our implementation & scaling characteristics:

The **Map** method just **extracts** the **ASIN and the ReviewText** from the line, which is in JSON format. We **loop over each word** of the ReviewText, and **if the word is a positive or negative word, it is written to the context**. Here is an example for a positive word: context.write(new Text("P" + asin), one);. So for each ASIN, there are two keys: P<ASIN> and N<ASIN>.

The **Reduce** method just **sums** the **counts** up **for each ASIN**.

After the MapReduce job is completed, the we take a look at the **result files in HDFS**. We create a **hashmap**. The **key is the ASIN, the value is an Integer array with a length of two**. It contains the number of positive and negative words for this ASIN. Then we just calculate the sentiment and positivity score for each ASIN and print them.

Number of invocations of the Map and Reduce methods:

For the electronics dataset:
The json file has 1,689,188 lines. For each line, the map method is called one time. Since there were 11 map tasks (which could be run in parallel theoretically), each of those tasks would invoke the Map method about 153,562 times.
There are 63,001 distinct ASINs in the dataset. So our Map would have 126,002 keys (each ASIN has two keys, one for positive and one for negative words). For each of them, there is at least one call of the Reduce function.

Runtime spent in the different phases of the process:

For the electronics dataset the results were:

Total time spent by all map tasks	258,676ms
Total time spent by all reduce tasks	16,643ms.

And the average time for each of the tasks of a MapReduce job:

Average Map Time	23sec
Average Shuffle Time	15sec
Average Merge Time	0sec
Average Reduce Time	0sec

But there were 11 Map tasks, which were run sequentially. The average map time was 23 seconds. There was only one reduce task, which took 15 seconds to complete (actually, the shuffle task took almost all the time of the reduce task).

What speedup would you expect from distributing the job among more machines?

Almost linear speedup, since MapReduce scales very well, and our scenario allows for an efficiently parallelizable MapReduce implementation. (Creating a map with ‘ones’ for each positive/negative word, and then summing them up in the reduce task).

Sentiment scores of the first 10 products of each category:

Electronics

B004ZC2JFS: 0.4883721
B00IVPU5BK: 0.5344828
B00HL7Z46K: 0.6585366
B001250KX4: 0.3488372
B003CK10DG: 0.27034122
B0014HS938: 0.50163937
B002NX7M9O: 0.52380955
B0065XBGCC: 0.5
B00022PUDI: 0.14857143
B004RKVWQK: 0.6103896

Video Games

B000AQA9UA: 0.57990867
B0027IS82K: 0.4678899
B00006LEM9: 0.2488263
B000067A3M: 0.14204545
B000P297ES: 0.29149798
B00BMFIXT2: 0.25811437
B00006LEM8: 0.04901961
B000P297F2: 0.019741321
B00009V3NB: 0.11627907
B0008F6K06: 0.0

Baby

B003VNKLIY: 0.49387756
B0042TO3NK: 0.34020618
B002VA2UFA: 0.47058824
B003YVHXAW: 0.49056605
B002YMB5JM: 0.6363636
B0017T4PXG: 0.73913044
B002GP76K4: 0.33668342
B000P1U9VE: 0.4857143
B001RLUS5S: 0.42857143
B00DHINGB2: 0.456446

2. MovieLens Dataset Analysis with Hive

Remarks:

Due to commas in the movie titles, we use SerDe instead of the default delimiter:
https://tuwel.tuwien.ac.at/mod/forum/discuss.php?d=102893

The limitation of SerDe is, that all columns are handled as Strings, even though the datatypes of the columns are set to something different (e.g. INT) in the **create.hql** file. But for our task, this should not make a difference, since this did not seem to make a difference for the calculations (except for the UNIX timestamp, which had to be explicitly casted to INT) of aggregators like AVG (perhaps, the entries are auto-casted to a fitting datatype for the aggregator).
https://cwiki.apache.org/confluence/display/Hive/CSV+Serde

Query Results:

1) How many movie ratings are there in total in the dataset?

- 26024289

2) How many movies in the dataset belong to the "Horror" genre?

- 4448

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

- sci-fi
- atmospheric
- action
- comedy
- based on a book
- surreal
- twist ending
- funny
- BD-R
- classic

4) Which 10 movies were the most controversial in 2010 (i.e., had the highest variance in ratings between 2010/01/01 and 2010/12/31)?

- Ethan Mao (2004)
- Sudden Fear (1952)
- Delgo (2008)
- I Got the Hook Up (1998)
- Unidentified Flying Oddball (a.k.a. Spaceman and King Arthur, The) (a.k.a. Spaceman in King Arthur's Court, A) (1979)
- Clifford's Really Big Movie (2004)
- Lady Death (2004)

- Ranma ½: Big Trouble in Nekonron, China (Ranma ½: Chûgoku Nekonron daikessen! Okite yaburi no gekitô hen) (1991)
- Kids of the Round Table (1995)
- Angels of the Universe (Englar alheimsins) (2000)

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

- Star Wars: Episode IV - A New Hope (1977) 9204
- Pulp Fiction (1994) 4454
- Inception (2010) 3972
- Shawshank Redemption, The (1994) 3355
- Fight Club (1999) 3331
- Matrix, The (1999) 3116
- Interstellar (2014) 2918
- Forrest Gump (1994) 2754
- Memento (2000) 2172
- Eternal Sunshine of the Spotless Mind (2004) 2087

6) Which 15 movies (titles) have been most frequently tagged with the label "sci-fi"?

- Star Wars: Episode IV - A New Hope (1977) 1035
- Matrix, The (1999) 261
- Interstellar (2014) 186
- Inception (2010) 157
- Blade Runner (1982) 132
- Avatar (2009) 132
- Ex Machina (2015) 131
- Star Wars: Episode V - The Empire Strikes Back (1980) 114
- Alien (1979) 107
- 2001: A Space Odyssey (1968) 104
- Fifth Element, The (1997) 103
- District 9 (2009) 98
- Star Wars: Episode VI - Return of the Jedi (1983) 91
- Edge of Tomorrow (2014) 91
- The Martian (2015) 83

7) Which are the 10 best-rated movies (on average; list titles) with more than 500 ratings?

- Planet Earth (2006)
- Shawshank Redemption, The (1994)
- Godfather, The (1972)
- Usual Suspects, The (1995)
- Schindler's List (1993)
- Godfather: Part II, The (1974)
- Seven Samurai (Shichinin no samurai) (1954)
- Rear Window (1954)
- 12 Angry Men (1957)
- Fight Club (1999)

8) Which are 10 highest-rated "Drama" movies with more than 10 ratings?

- Shawshank Redemption, The (1994)
- Band of Brothers (2001)
- Godfather, The (1972)
- Schindler's List (1993)
- Godfather: Part II, The (1974)
- Seven Samurai (Shichinin no samurai) (1954)
- 12 Angry Men (1957)
- Fight Club (1999)
- One Flew Over the Cuckoo's Nest (1975)
- Over the Garden Wall (2013)

9) What are the 15 most relevant genome tags for the movie "Four rooms" (movieId=18)?

- off-beat comedy
- hotel
- storytelling
- weird
- multiple storylines
- stylish
- tarantino
- original
- great ending
- dark humor
- twists & turns
- dialogue
- great acting
- absurd
- comedy

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

- Third Man, The (1949)
- Johnny Guitar (1954)
- Before Sunrise (1995)
- Before Sunset (2004)
- Before Midnight (2013)
- Woman in Gold (2015)
- Night Porter, The (Portiere di notte, Il) (1974)
- Amadeus (1984)
- Illusionist, The (2006)
- Foreign Affair, A (1948)

Behind the scenes:

Apache Hive provides us with an **SQL-like query language** for unstructured Big Data. This way, we **do not have to write MapReduce jobs in the underlying Java API**. Instead, it

seems like the unstructured Data has a relational structure. (Other execution engines beside MR are taz and Spark).
So by executing a HQL query, the **query is translated to a MapReduce job**, and submitted to Hadoop for execution.

Since MapReduce allows for almost linear speedup, running the HQL queries on a real cluster in parallel would **speed up the query time almost linear to the number of processors** available in our cluster (capped by the amount of Map/Reduce jobs required).

On a side note, you could inspect the state of the MapReduce job for each query (of the past day or so) on the URL **quickstart.cloudera:8088** inside the Virtual Machine.
On this page, you could see how many Map jobs and how many Reduce Jobs were required for task completion as well as the progress of the task.
Notably, there were usually more Reduce than Map jobs (typically around 3 Map jobs vs. 11 Reduce jobs), but the Reduce Jobs took only a small fraction of the time.

3. Spark Movie Recommender

Implementation

The movies.csv files had commas in the movie titles so we had to parse the files at first to get a clean looking output at the end. To achieve this we copied the files into tmp files during parsing that delete themselves at the end of the programm, to leave the source files untouched.
The next step was splitting the datasets into test and training sets and fusing them with our own ratings.
Then we repeatedly trained the set to find out the optimal attributes for the final recommendation. It happened to be at **rank = 15**, **iterations = 13** and **lambda = 0.04**.

At the end we just need to put the results on the screen and we can see our recommendations!

Ratings and recommendations of the group members

Roman, Ratings:

- Toy Story (1995) 5.0
- Jumanji (1995) 5.0
- Pocahontas (1995) 3.0
- Aladdin (1992) 4.0
- Harry Potter and the Deathly Hallows: Part 1 (2010) 5.0
- Green Hornet, The (2011) 5.0
- Winnie the Pooh and Tigger Too (1974) 1.0
- Kung Fu Panda 2 (2011) 5.0
- X-Men: First Class (2011) 4.0
- Harry Potter and the Deathly Hallows: Part 2 (2011) 5.0
- Smurfs, The (2011) 2.0
- Avengers, The (2012) 5.0
- Sixth Sense, The (1999) 3.0
- Three Musketeers, The (2011) 3.0
- John Carter (2012) 5.0
- Godfather, The (1972) 3.0
- Dictator, The (2012) 5.0
- Silence of the Lambs, The (1991) 3.0
- Brave (2012) 5.0
- Planes (2013) 2.0
- Zombeavers (2014) 2.0
- Haunt (2013) 1.0
- Interview with the Vampire: The Vampire Chronicles (1994) 1.0
- Jurassic Park (1993) 2.0
- Day of the Dead (1985) 1.0

Results:

Uno: The Movie	Comedy
Under the Hawthorn Tree (2010)	Drama, Romance
Tintin and the Lake of Sharks (1972)	Adventure,Animation,Children,Mystery
Lost in a Harem (1944)	Comedy
The Challengers (1990)	Children, Drama
Memories (2013)	Crime, Thriller
Unfair Competition (Concorrenza sleale) (2001)	Drama, War
Shock Head Soul (2011)	Documentary
The Monkey King the Legend Begins	Action, Adventure, Fantasy
Happy Family (2010)	Comedy
The Paper Brigade (1997)	Adventure, Children, Comedy
"Il ricco, il povero e il maggiordomo (2014)"	Comedy
The Brainwashing of My Dad (2015)	Documentary
Big Easy Express (2012)	Documentary, Musical
Kick (2009)	Action,Comedy,Romance,Thriller
If Tomorrow Comes (1986)	Crime, Drama, Mystery
"Brasher Doubloon, The (1947)"	Crime,Drama,Film-Noir,Mystery
Leader (2010)	Drama, Romance
Journey to the Center of the Earth (1989)	Action, Children, Fantasy, Sci-Fi
Manoman (2015)	Animation, Drama

Max, Ratings:

• Prestige, The (2006)	5.0
• Thing, The (1982)	5.0
• Memento (2000)	5.0
• From Dusk Till Dawn (1996)	5.0
• Star Wars: Episode II - Attack of the Clones (2002)	2.0
• Shawshank Redemption The (1994)	5.0
• Shining, The (1980)	3.0
• The Similars (2015)	3.0
• Devil (2010)	2.0
• [REC] (2007)	5.0
• [REC]² (2009)	4.0
• Quarantine 2: Terminal (2011)	1.0
• Book of Eli, The (2010)	4.0
• Sucker Punch (2011)	1.0
• Hot Fuzz (2007)	5.0
• Ender's Game (2013)	4.0
• Ghostbusters (a.k.a. Ghost Busters) (1984)	5.0
• The Dark Knight (2011)	5.0
• 1408 (2007)	4.0
• Godzilla (2014)	1.0
• Station, The (Blutgletscher) (2013)	4.0

- Cheap Thrills (2013) 4.0
- 2 Fast 2 Furious (Fast and the Furious 2, The) (2003) 2.0
- Scary Movie 5 (Scary MoVie) (2013) 2.0
- Pacific Rim (2013) 2.0

Results:

The Thorn (1971)	Comedy
Pizza (2012)	Comedy,Horror,Romance,Thriller
Kaaka Muttai (2015)	(no genres listed)
Hampstead (2017)	Comedy, Romance
AM1200 (2008)	Horror
Emo Philips Live (1987)	Comedy
Jimi Plays Monterey (1986)	(no genres listed)
Return to Source: The Philosophy of The Matrix (2004)	Documentary
The Great Piggy Bank Robbery (1946)	Animation, Children, Comedy
The Garden of Afflictions 2017	(no genres listed)
The Wearing of the Grin (1951)	Animation
Trailer Park Boys - Live in F**kin' Dublin (2014)	Comedy
101 Rent Boys (2000)	Documentary
"Rise & Fall of ECW, The (2004)"	Documentary
Shivering Trunks	(no genres listed)
The Spirit of Christmas (1995)	Animation, Comedy
Athadu (2005)	Action, Thriller
Christmas on Salvation Street (2014)	(no genres listed)
The Take (2009)	(no genres listed)
Love Is Blind (2013)	Drama, Romance

Wolfi, Ratings:

- Harry Potter and the Deathly Hallows: Part 2 (2011) 5.0
- Harry Potter and the Deathly Hallows: Part 1 (2010) 5.0
- Lion King, The (1994) 5.0
- Toy Story (1995) 5.0
- Cars (2006) 5.0
- Robin Hood (1972) 5.0
- Monsters, Inc. (2001) 5.0
- Bambi (1942) 2.0
- Despicable Me (2010) 5.0
- Rush (2013) 4.0
- Senna (2010) 4.0
- Star Wars: Episode III - Revenge of the Sith (2005) 5.0
- Star Wars: Episode VII - The Force Awakens (2015) 5.0
- Borat: Cultural Learnings of America for Make Benefit Glorious Nation of Kazakhstan (2006) 5.0

- Anabelle (2014) 2.0
- Inglorious Basterds (2009) 5.0
- Django Unchained (2012) 4.0
- Forrest Gump (1994) 5.0
- Clockwork Orange (1971) 3.0
- American Sniper (2014) 2.0
- Suicide Squad (2016) 3.0
- American Pie (1999) 1.0
- Scary Movie (2000) 2.0
- Avengers, The (2012) 2.0
- Click (2006) 1.0

Results:

On Body and Soul (2017)	Drama, Romance
The Thorn (1971)	Comedy
The Wearing of the Grin (1951)	Animation
Lewis Black: Red, White & Screwed (2006)	Comedy
The Great Piggy Bank Robbery (1946)	Animation, Children, Comedy
The Spirit of Christmas (1995)	Animation, Comedy
Advanced Style (2014)	Comedy, Documentary, Drama
Lewis Black: Black on Broadway (2004)	Comedy
Barking at the Stars (1998)	Action,Comedy,Romance,Sci-Fi,Thriller
Smashing Pumpkins: Vieuphoria (1994)	Documentary, Musical
The Spirit of Christmas (1992)	Animation, Comedy, Thriller
Final (2001)	Drama, Sci-Fi, Thriller
Lush Life (1993)	Drama, Musical
Life Even Looks Like a Party (2009)	Documentary
Lewis Black: In God We Rust (2012)	Comedy
The Garden of Afflictions 2017	(no genres listed)
Deli Man (2015)	Documentary
I'll Take You There (1999)	Comedy, Drama, Romance
Neue Vahr Süd (2010)	Comedy
Doggiewoggiez! Poochiewoochiez! (2012)	Comedy

4. Appendix

4.1 MapReduce

Create Shared Folder for VirtualBox:

1) create folder in host OS
2) in VirtualBox:
    Devices -> Shared Folders -> Shared Folders Settings...
    Shared Folders -> Add (Icon, to the right)
        Folder Path: Path to previously created Folder
        Folder Name: Name you want the folder to have (**FOLDER_NAME**)

[x] Make permanent: because why not

3) in guest OS:

```
mkdir /media/shared
mount -t vboxsf -o uid=$UID,gid=$(id -g) FOLDER_NAME /media/shared
```

where **FOLDER_NAME** is the name of the shared folder in virtual box.

Problem: Hadoop Permission Denied:
Prepend `sudo -u hdfs` to the command (execute it as user hdfs)
https://stackoverflow.com/questions/22676593/why-does-hadoop-fs-mkdir-fail-with-permission-denied

Search for Maven Dependencies: (and how to refer to them in pom.xml)
https://mvnrepository.com/search?q=apache+hadoop

Referenced Maven Packages:

- `hadoop-common`
- `hadoop-client`

"Upload" pos-words.txt & neg-words.txt to HDFS

```
hadoop fs -copyFromLocal pos-words.txt .
hadoop fs -copyFromLocal neg-words.txt .
```

View the files in HDFS:
`hdfs dfs -ls` (or `hadoop fs -ls`)

Run our MR.jar:
For execution, cannot easily be run directly from Eclipse as it has to access files in the HDFS.
Thus, we had to export our project as a runnable JAR file and then run it via the command:

```
hadoop jar MR.jar /amazon/reviews_Baby_5SMALL.json
/amazon/outputBabySMALL19 -pos pos-words.txt -neg neg-words.txt
```

4.2 MovieLens

Hive / Beeline
In the beginning, we only started the services mentioned in the Assignment PDF (i.e. HDFS, YARN, Hive, Hue) and tried both Hue and the Hive CLI to issue the queries.
Then, very simple queries took either a long time (e.g. 60s for `create/use database`) or did not terminate at all (`create table …` was stopped after over 3 hours).
After that, we tried out if beeline (`beeline -u "jdbc:hive2://"`) would make any difference. It complained about Zookeeper not running, so we started Zookeeper.
After that, beeline performed fast enough to be usable. Interestingly enough, both Hue and Hive worked as well after starting Zookeeper, so that we resorted back to the more comfortable Web UI of Hue.

Preventing beeline from crying verbosely about Access Permissions to Logs:
Make it run as somebody that has permissions (e.g. root… great idea)

```
sudo beeline -u "jdbc:hive:2://"
```

Loading the Datasets from HDFS into the Tables:
Make it run as somebody that has permissions (e.g. hdfs)

```
sudo -u hdfs hive
> LOAD INPATH …
```

4.3 Spark

Problems
At first we started a Scala project in IntelliJ and tried to compile some Spark library imports. Then the first problem arose when our Scala version was too new for Spark and we had to downgrade it.
After a few hours of programming we tried our working program inside the Spark shell, but first we needed a file called winutils.exe. It still didn't work for some time until we found out that we needed the 64 bit version.
The last problem was that if we extended the Scala file from App, it wasn't serializable.

To be able to use the big dataset we increased the spark executor memory to 8 GB.

Starting the program from the command line
Starting Spark:

> <path>/spark-shell -i <path>/MovieRecommendations.scala

Starting the program in sparc

> MovieRecommendations.main(null)