检测Spark应用中的缓存相关缺陷

李慧，王栋，黄天泽，高钰，窦文生，许利杰，王伟，魏峻，钟华

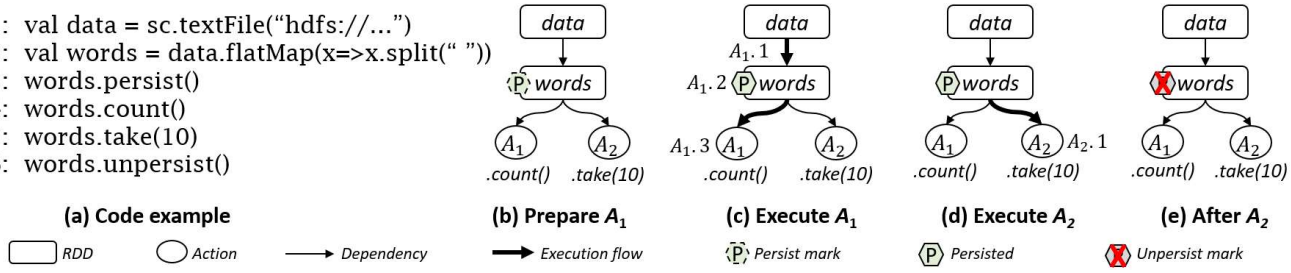
Detecting Cache-Related Bugs in Spark Applications, ISSTA 2020

联系方式：王栋，wangdong18@otcaix.iscas.ac.cn，http://www.tcse.cn/~wangdong18

Cache-Related Bugs in Spark Applications

The execution process and cache mechanism in Spark

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: words.persist()
4: words.count()
5: words.take(10)
6: words.unpersist()
```



Cache-related bugs

- Missing persist, lagging persist and premature unpersist bugs can cause the duplicated computation.
- Unnecessary persist, missing unpersist and lagging unpersist bugs can cause the waste of memory.

a) Missing persist

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: ~~words.persist()~~
4: words.count()
5: words.take(10)
```

b) Lagging persist

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: ~~words.persist()~~
4: words.count()
5: words.persist()
6: words.take(10)
```

c) Premature unpersist

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: words.persist()
4: words.count()
5: words.unpersist()
6: words.take(10)
```

d) Unnecessary persist

```
1: val data = sc.textFile("hdfs://...")
2: data.persist()
3: val words = data.flatMap(x=>x.split(" "))
4: words.persist()
5: words.count()
6: words.take(10)
```

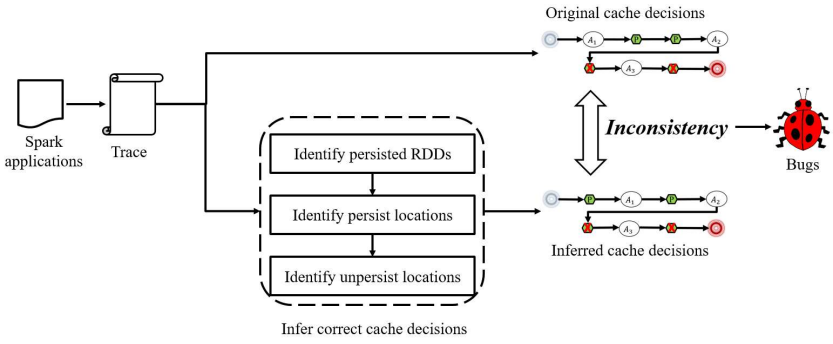
e) Missing unpersist

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: words.persist()
4: words.count()
5: words.take(10)
6: ~~words.unpersist()~~
...
```

f) Lagging unpersist

```
1: val data = sc.textFile("hdfs://...")
2: val words = data.flatMap(x=>x.split(" "))
3: words.persist()
4: words.count()
5: words.take(10)
6: ~~words.unpersist()~~
...
20: words.unpersist()
```

CacheCheck

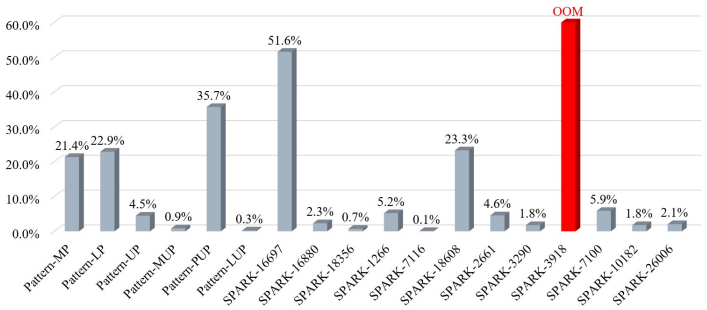


We present CacheCheck to detect cache-related bugs in Spark applications.

- Run the application and collect traces.
- Extract original cache decisions.
- Infer correct cache decisions on should-be-persisted RDDs.
- Detect cache-related bugs by comparing the original cache decisions and the inferred cache decisions.

Evaluation

Performance slowdown



New bug detection

App	MP	LP	UP	MUP	LUP	PUP	Total	Confirmed	Fixed
GraphX	1	0	6	1	0	0	8	7	0
MLlib	24	1	10	2	12	2	51	43	27
Spark SQL	4	0	0	0	0	0	4	3	1
MCL	4	0	0	0	0	0	4	0	0
Betweenness	0	0	1	0	0	0	1	0	0
t-SNE	1	0	1	2	0	0	4	0	0
Total	34	1	18	5	12	2	72	53	28