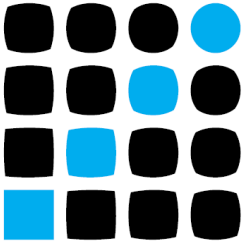


adaptTo()

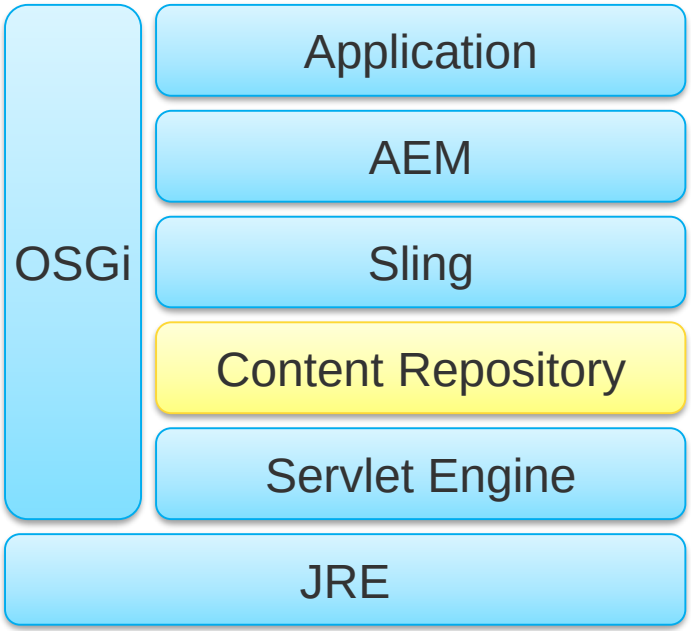
APACHE SLING & FRIENDS TECH MEETUP

BERLIN, 26-28 SEPTEMBER 2016

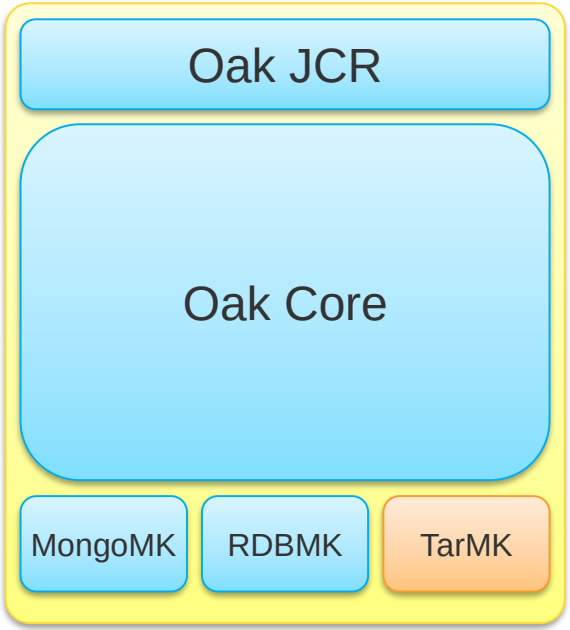
Into the Tar Pit: A TarMK Deep Dive

Michael Dürig, Adobe Research

TarMK


OSGi
Application
AEM
Sling
Content Repository
Servlet Engine
JRE

TarMK



TarMK

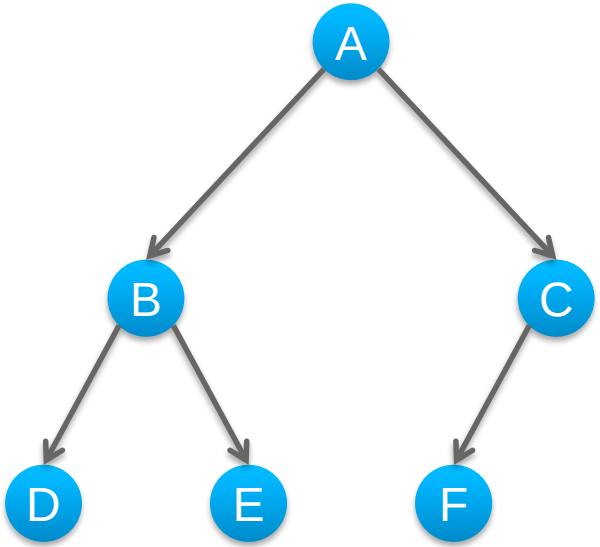
- Embedded Database
  - Hierarchical
  - Fast / Small
  - Limited scalability
  - MVCC / Append only

Agenda

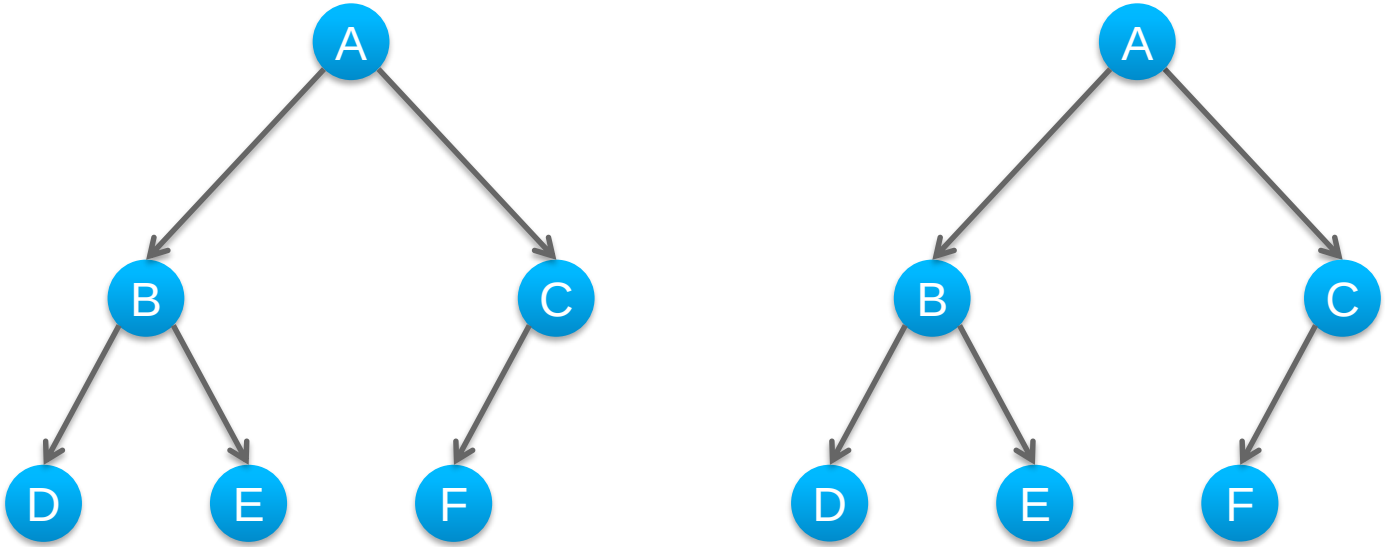
- MVCC Persistence
- Revisions, Recovery, Rollback
- Garbage Collection
- Upcoming Improvements

MVCC Persistence

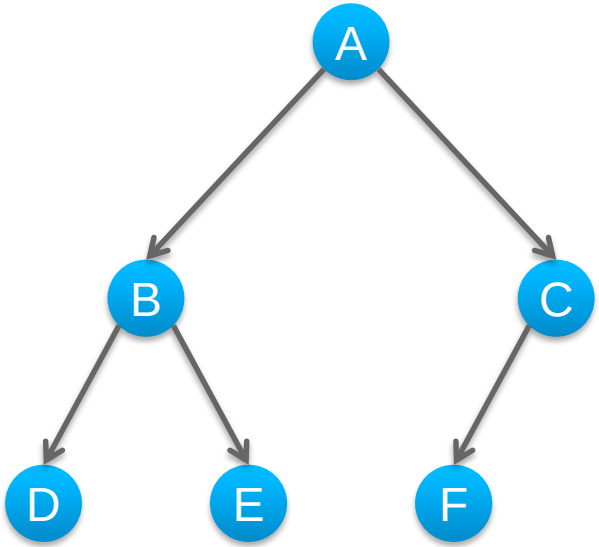
Updating Trees

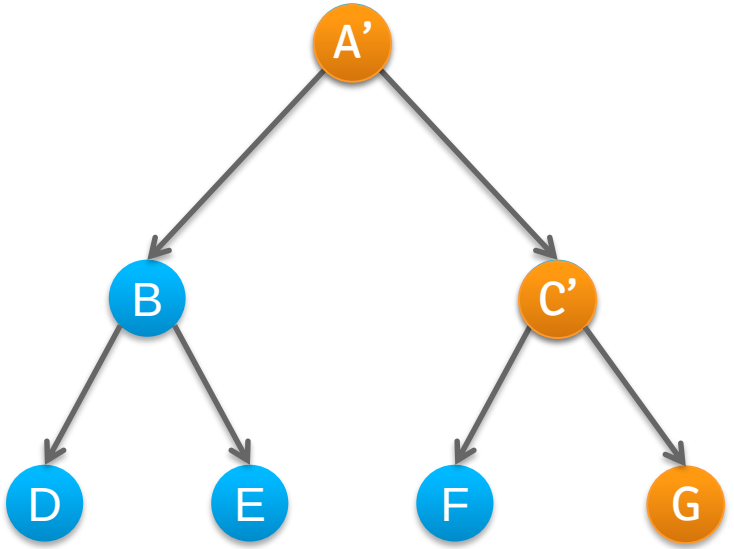


Updating Trees

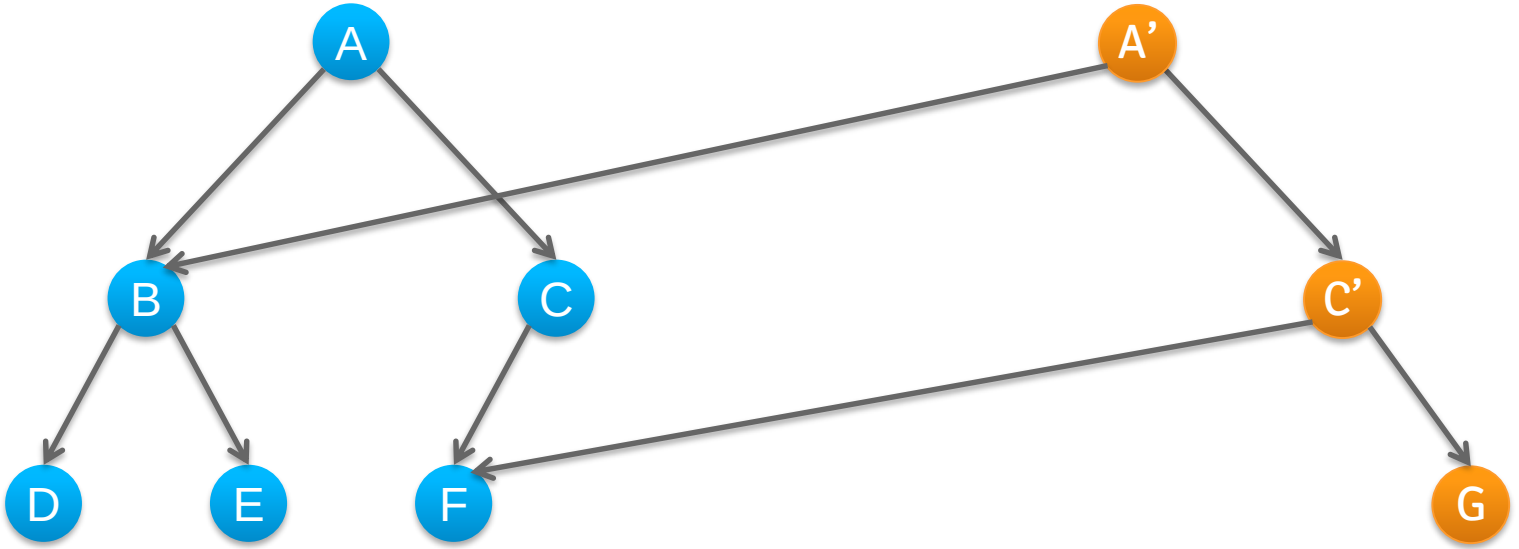


Updating Trees



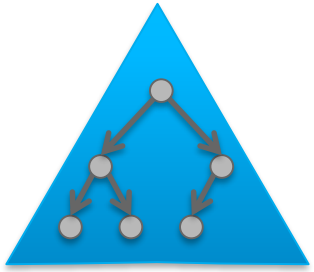


Updating Trees

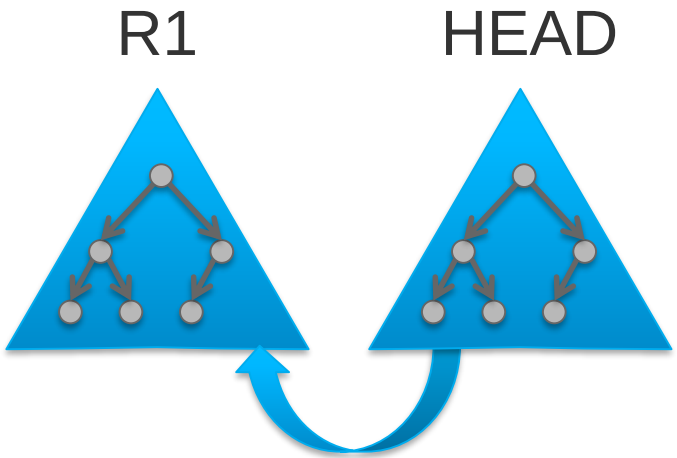


Revisions

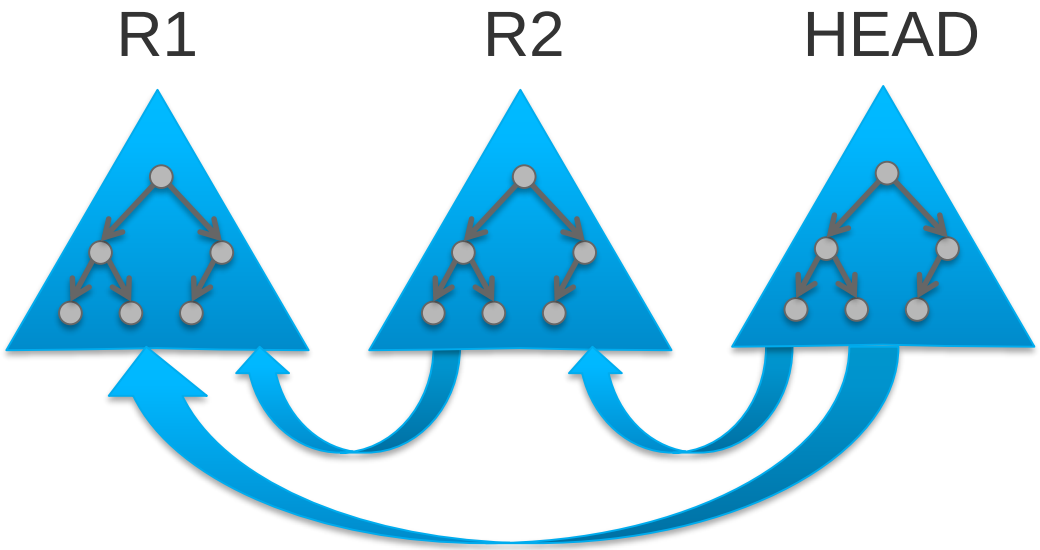
HEAD



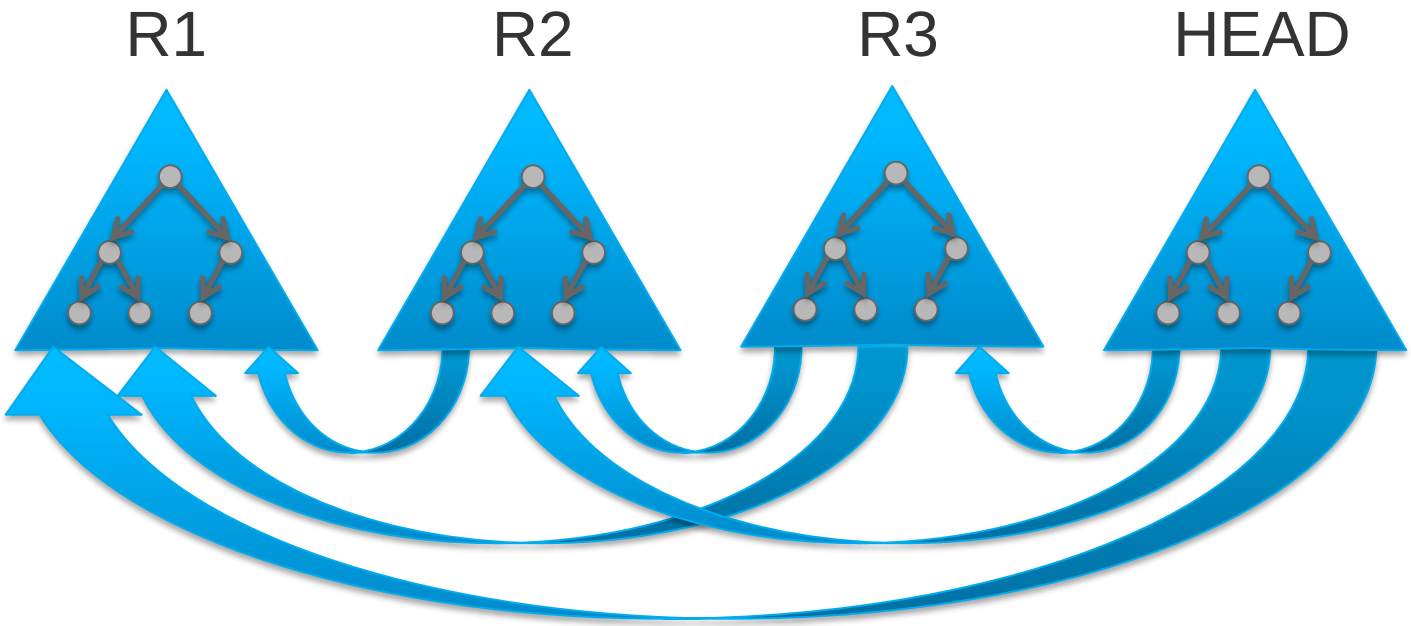
Revisions



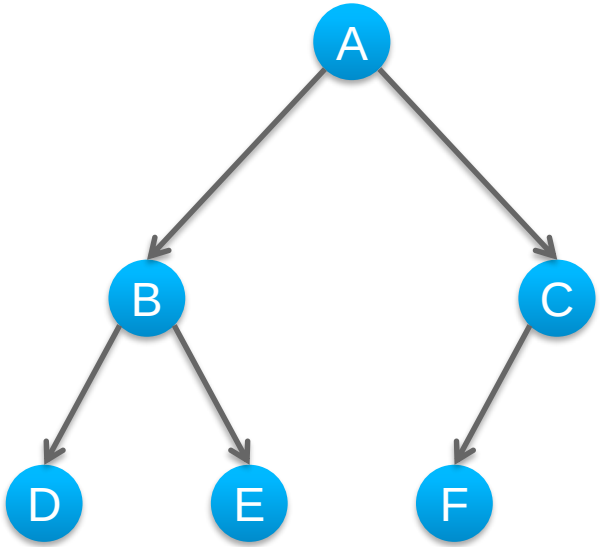
Revisions



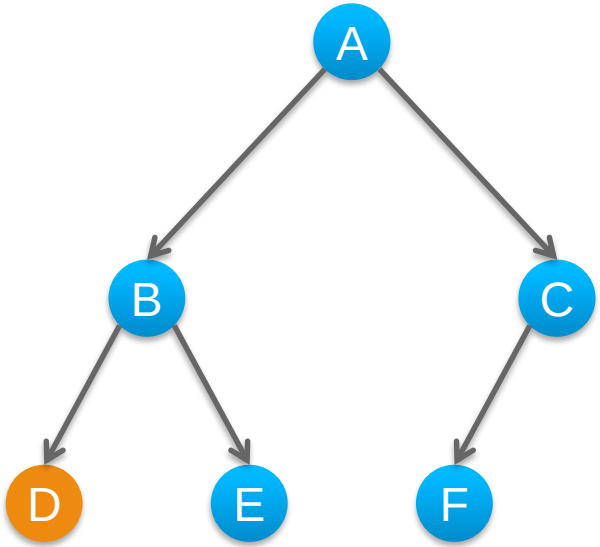
Revisions



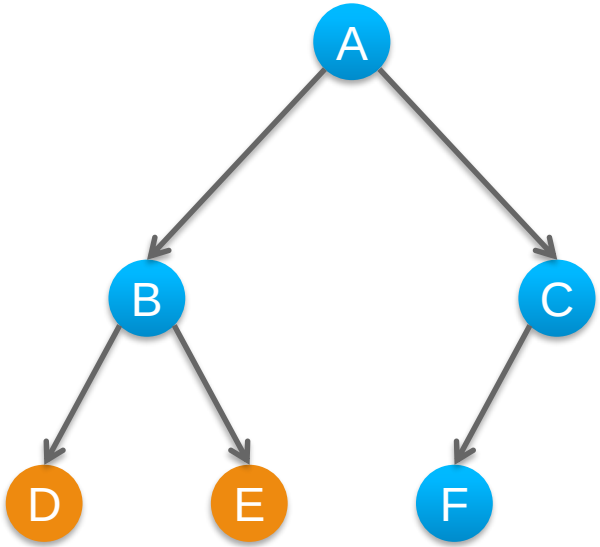
Persisting Revisions



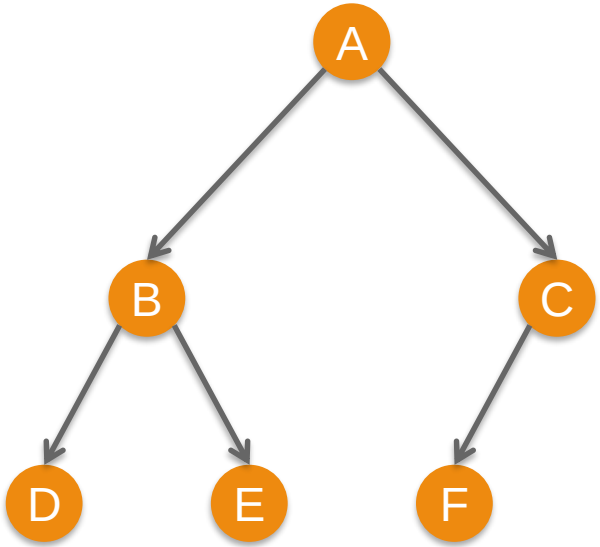
Persisting Revisions



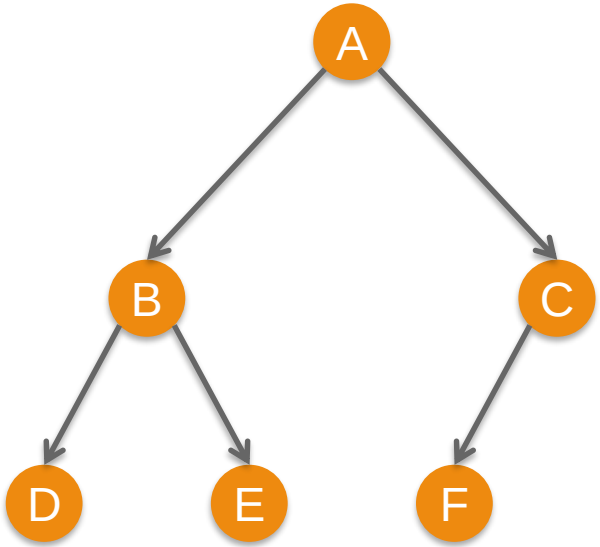
Persisting Revisions

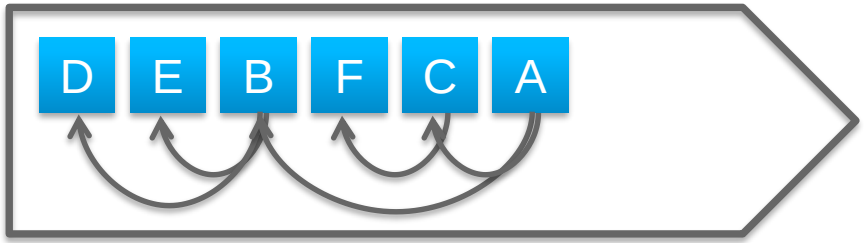


Persisting Revisions

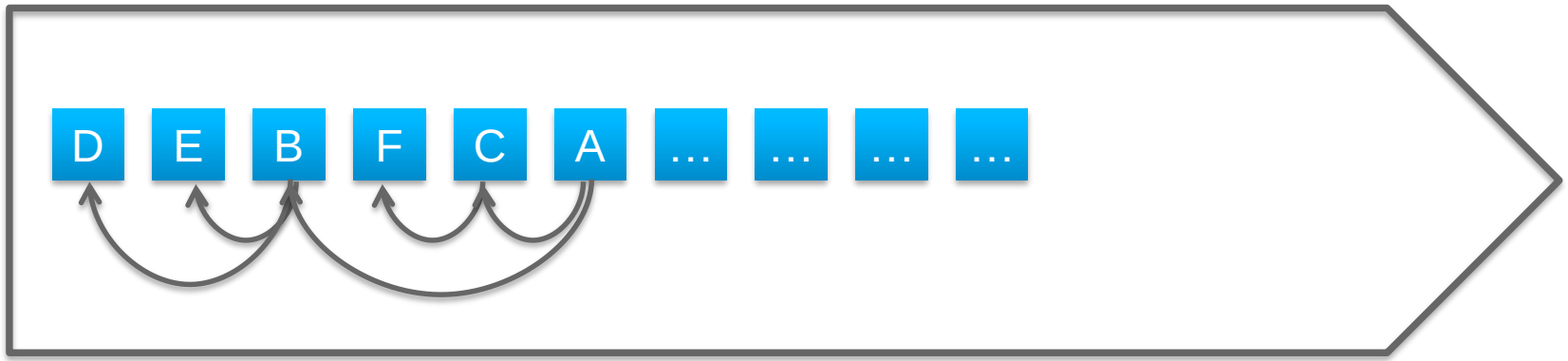


Persisting Revisions



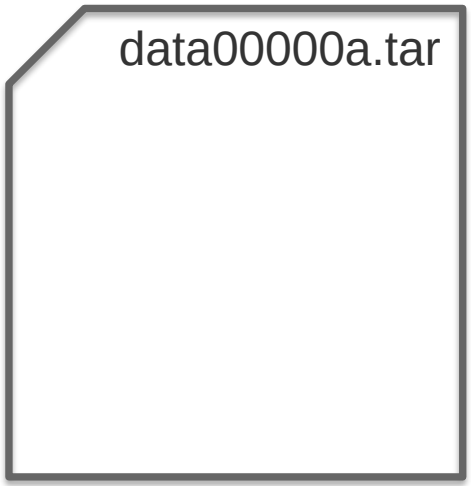


Records and Segments

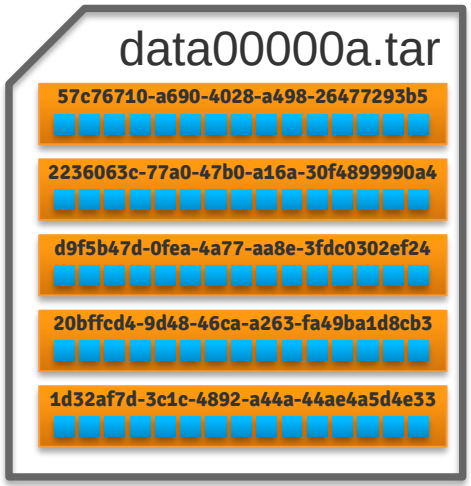


Records and Segments

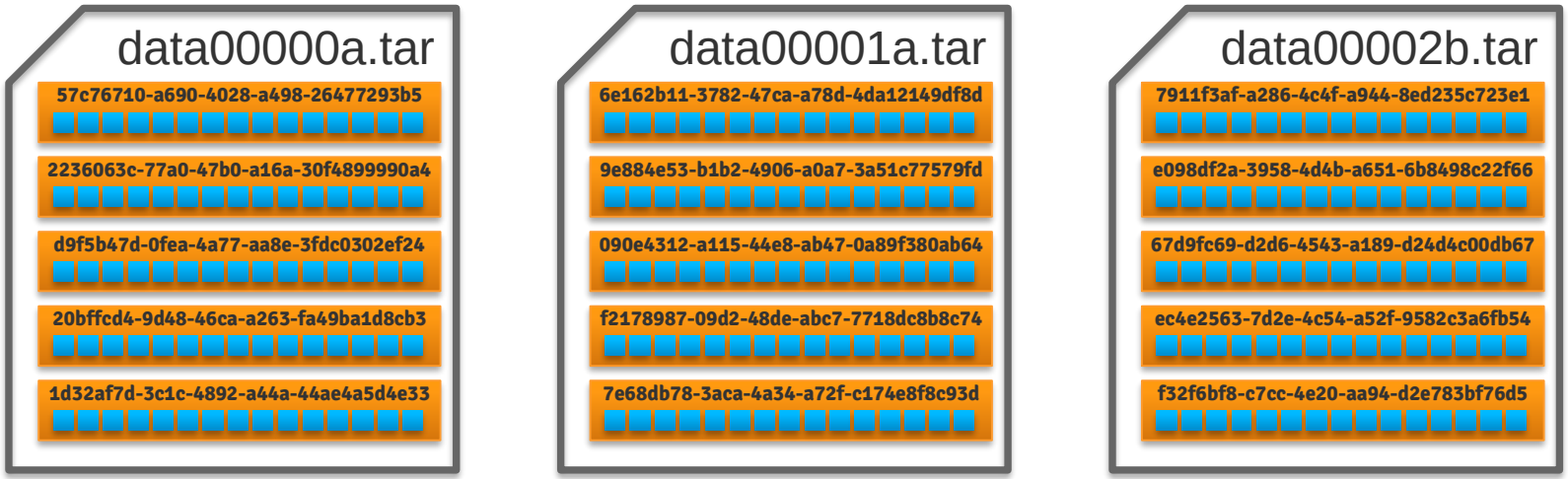
Tar Files



Tar Files


data00000a.tar
57c76710-a690-4028-a498-26477293b5
2236063c-77a0-47b0-a16a-30f4899990a4
d9f5b47d-0fea-4a77-aa8e-3fdc0302ef24
20bffcd4-9d48-46ca-a263-fa49ba1d8cb3
1d32af7d-3c1c-4892-a44a-44ae4a5d4e33

Tar Files


data00000a.tar
57c76710-a690-4028-a498-26477293b5
2236063c-77a0-47b0-a16a-30f4899990a4
d9f5b47d-0fea-4a77-aa8e-3fdc0302ef24
20bffcd4-9d48-46ca-a263-fa49ba1d8cb3
1d32af7d-3c1c-4892-a44a-44ae4a5d4e33
data00001a.tar
6e162b11-3782-47ca-a78d-4da12149df8d
9e884e53-b1b2-4906-a0a7-3a51c77579fd
090e4312-a115-44e8-ab47-0a89f380ab64
f2178987-09d2-48de-abc7-7718dc8b8c74
7e68db78-3aca-4a34-a72f-c174e8f8c93d
data00002b.tar
7911f3af-a286-4c4f-a944-8ed235c723e1
e098df2a-3958-4d4b-a651-6b8498c22f66
67d9fc69-d2d6-4543-a189-d24d4c00db67
ec4e2563-7d2e-4c54-a52f-9582c3a6fb54
f32f6bf8-c7cc-4e20-aa94-d2e783bf76d5

Revisions, Recovery, Rollback

Recovery

```
$ ls segmentstore
256M Aug 16 17:09 data00000a.tar
256M Aug 16 17:09 data00001a.tar
256M Aug 16 17:09 data00002a.tar
256M Aug 16 17:09 data00003a.tar
256M Aug 16 17:09 data00004a.tar
256M Aug 16 17:09 data00005a.tar
256M Aug 16 17:09 data00006a.tar
231M Aug 16 17:12 data00007a.tar
231M Aug 16 17:09 data00007a.tar.bak
256M Aug 16 17:12 data00008a.tar
182M Aug 16 17:12 data00009a.tar
147B Aug 16 17:09 journal.log
```

Recovery

```
$ ls segmentstore
256M Aug 16 17:09 data00000a.tar
256M Aug 16 17:09 d      1a.tar
256M Aug 16      .tar
```

```
$ tar -tvf data0000a.tar
 69644 Aug 16 17:09 0686c08d-e3f6-474e-bb32-874efca706e7.58dbce7b
262144 Aug 16 17:09 a3a57501-f30f-4986-b820-e166c50adaad.8d58d425
262144 Aug 16 17:09 8bf5c193-6458-4e29-b4f5-d0dbba5ca584.0a0ceeac
195080 Aug 16 17:09 2f6cbfdf-8d78-41d9-b507-89f47a143cab.47624a71
262144 Aug 16 17:09 7b8fd991-e894-49fb-b0e1-c193e5403755.da89a3db
262144 Aug 16 17:09 10d3ea6c-e62f-45d1-bd61-fab94db9cddd.da137b63
 25600 Aug 16 17:09 data00000a.tar.gph
 31232 Aug 16 17:09 data00000a.tar.idx
```

Recovery

```
$ ls segmentstore
256M Aug 16 17:09 data00000a.tar
256M Aug 16 17:09 data00001a.tar
256M Aug 16 17:09 data00002a.tar
256M Aug 16 17:09 data00003a.tar
256M Aug 16 17:09 data00004a.tar
256M Aug 16 17:09 data00005a.tar
256M Aug 16 17:09 data00006a.tar
231M Aug 16 17:12 data00007a.tar
231M Aug 16 17:09 data00007a.tar.bak
256M Aug 16 17:12 data00008a.tar
182M Aug 16 17:12 data00009a.tar
147B Aug 16 17:09 journal.log
```

Recovery

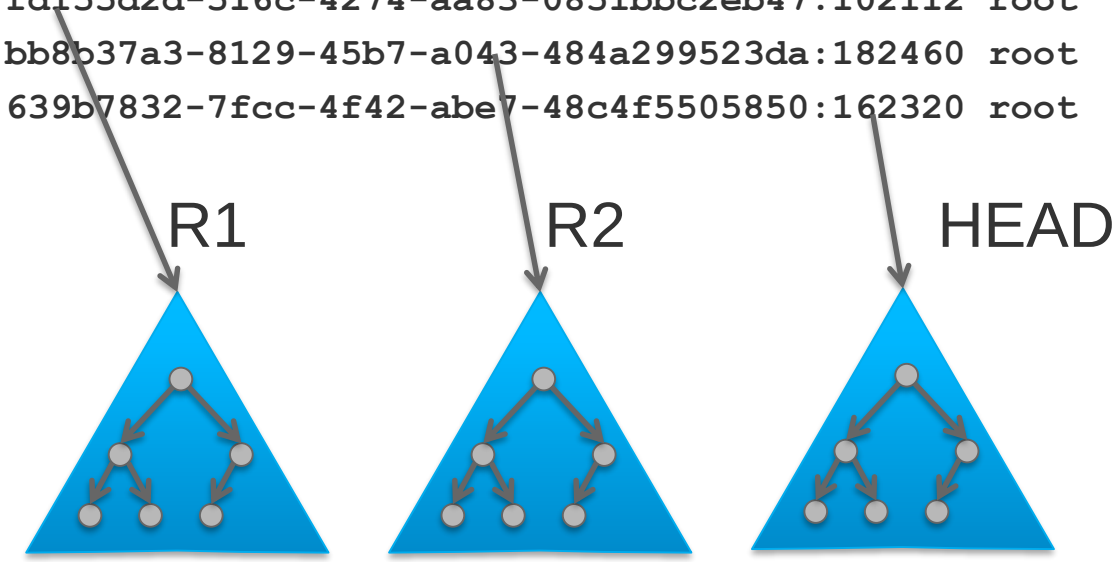
```
17:12:21.025 WARN Could not find a valid tar index in
[/segmentstore/data00007a.tar], recovering...

17:12:21.025 INFO Recovering segments from tar file
/segmentstore/data00007a.tar

17:12:21.739 INFO Backing up /segmentstore/data00007a.tar to
data00007a.tar.bak

17:12:21.739 INFO Regenerating tar file /segmentstore/data00007a.tar
```

Revisions

```
$ cat journal.log
fd155d2d-516c-4274-aa83-0851bbc2eb47:102112 root
bb8b37a3-8129-45b7-a043-484a299523da:182460 root
639b7832-7fcc-4f42-abe7-48c4f5505850:162320 root
```

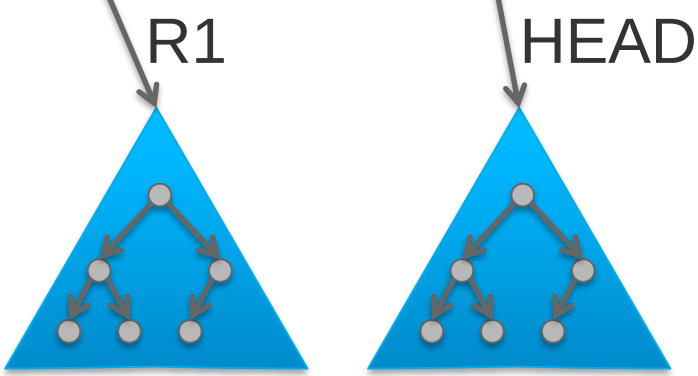
Revisions

```
$ cat journal.log
fd155d2d-516c-4274-aa83-0851bbc2eb47:102112 root
bb8b37a3-8129-45b7-a043-484a299523da:182460 root
639b7832-7fcc-4f42-abe7-48c4f5505850:162320 root
```



Revisions

```
$ cat journal.log
fd155d2d-516c-4274-aa83-0851bbc2eb47:102112 root
bb8b37a3-8129-45b7-a043-484a299523da:182460 root
639b7832-7fcc-4f42-abe7-48e4f550585b:162320 root
```



Rollback

```
$ java -jar oak-run-*.jar check

usage: check <options>
Option          Description
------          -----------
--bin [Long]    read the n first bytes from binary properties.
                -1 for all bytes. (default: 0)
--deep [Long]   enable deep consistency checking. An optional
                long specifies the number of seconds between
                progress notifications (default: 92233720368547758O7)
--journal       journal file (default: journal.log)
--path          path to the segment store (required)
```

Rollback

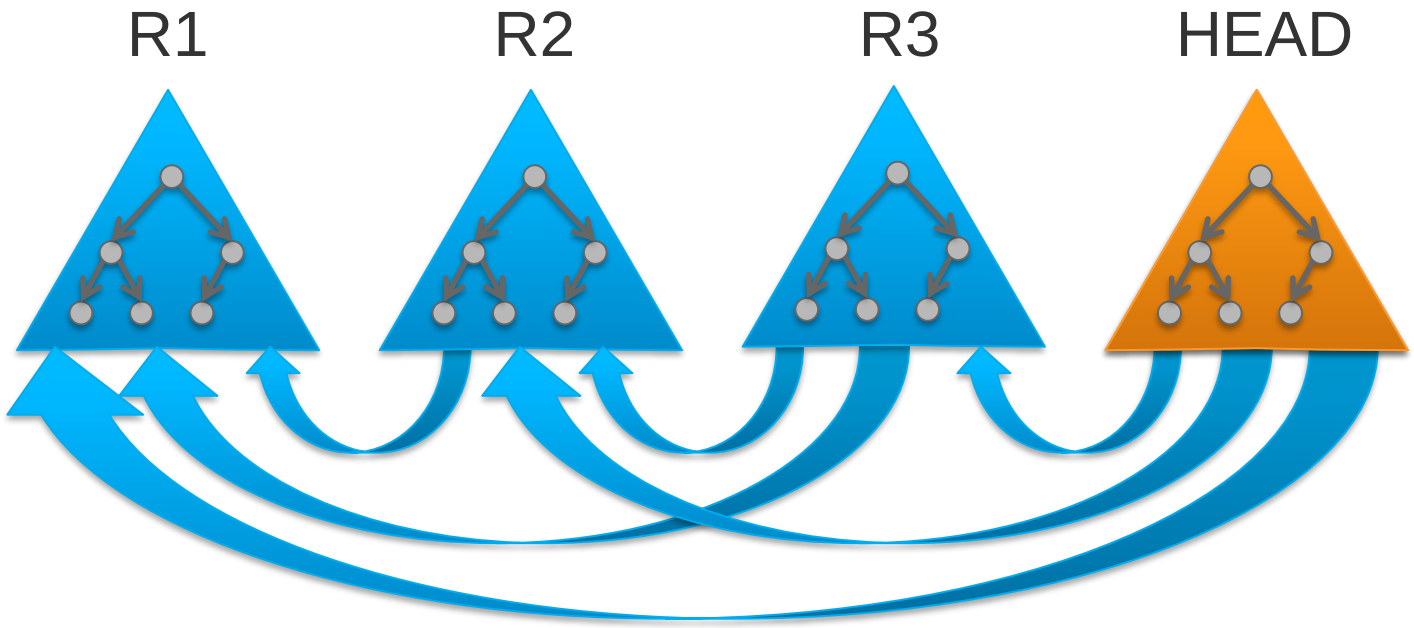
```
$ java -jar oak-run-*.jar check --deep --path /segmentstore

21:52:07.149 INFO  Searching for last good revision in journal.log

21:52:07.219 INFO  Checking revision 639b7832-7fcc-4f42-abe7-
48c4f5505850:162320

21:52:07.227 ERROR Segment not found: 639b7832-7fcc-4f42-abe7-
48c4f5505850. Creation date delta is 6 ms.

21:52:07.227 INFO  Error while traversing 639b7832-7fcc-4f42-abe7-
48c4f5505850:162320

21:52:07.228 INFO  Broken revision 639b7832-7fcc-4f42-abe7-
48c4f5505850:162320
```

Rollback

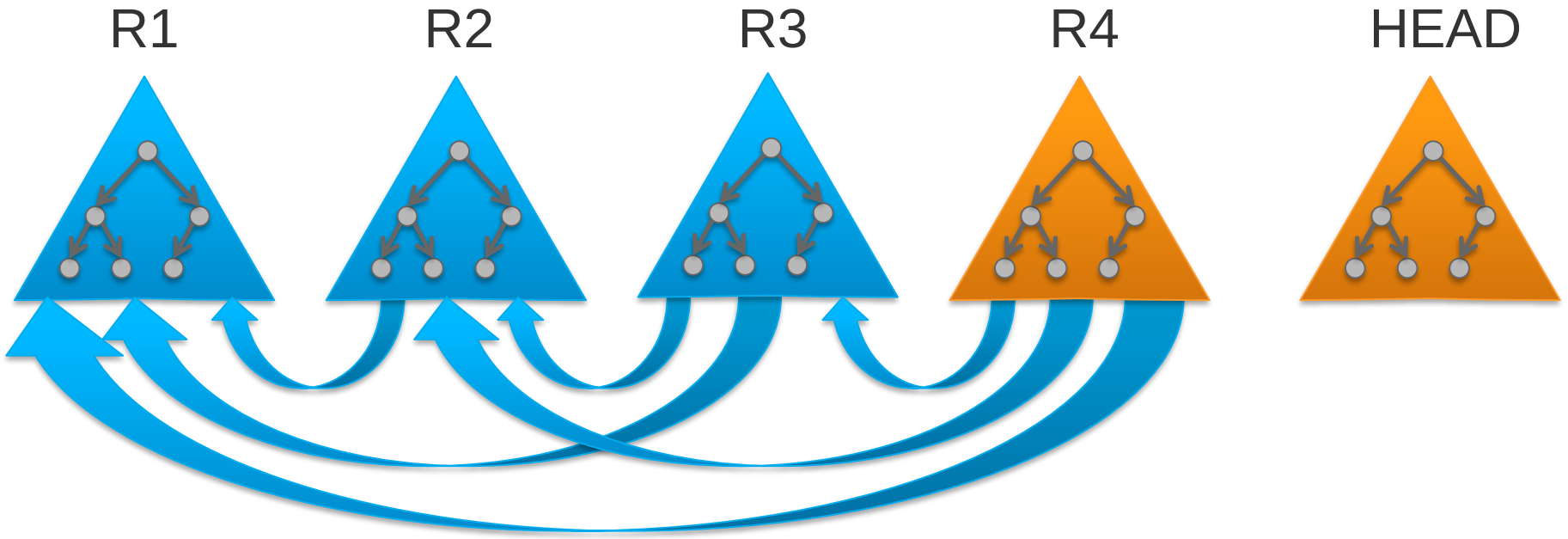
```
21:52:07.228 INFO  Checking revision bb8b37a3-8129-45b7-a043-
484a299523da:182460

21:52:07.228 INFO  Checking /

21:52:07.266 INFO  Traversed 88 nodes and 103 properties

21:52:07.266 INFO  Found latest good revision bb8b37a3-8129-45b7-a043-
484a299523da:182460

21:52:07.266 INFO  Searched through 2 revisions
```

Garbage Collection

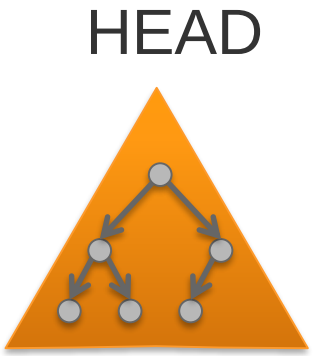
Offline Revisions GC



Offline Revisions GC



Offline Revisions GC



Offline Revision GC

```
$ java -jar oak-run-*.jar compact segmentstore

Compacting segmenstore
    before
        Tue Aug 16 17:09:05 CEST 2016, data00000a.tar
        Tue Aug 16 17:09:08 CEST 2016, data00001a.tar
        ...
    size 2.6 GB (2582279827 bytes)
    -> compacting
    -> cleaning up
    -> removed old file data00000a.tar
    -> removed old file data00001a.tar
    ...
```

Offline Revision GC

```
 -> writing new journal.log:
       3b632859-fafd-4113-a53a-335451933862:231132 root

    after
        Tue Aug 23 11:45:08 CEST 2016, data00000b.tar
        Tue Aug 23 11:45:08 CEST 2016, data00001b.tar
        ...
    size 546.4 MB (546363953 bytes)
    removed files [data00000a.tar, ...]
    added files [data00000b.tar, ...]
Compaction succeeded in 4.240 s (4s).
```

Online Revisions GC

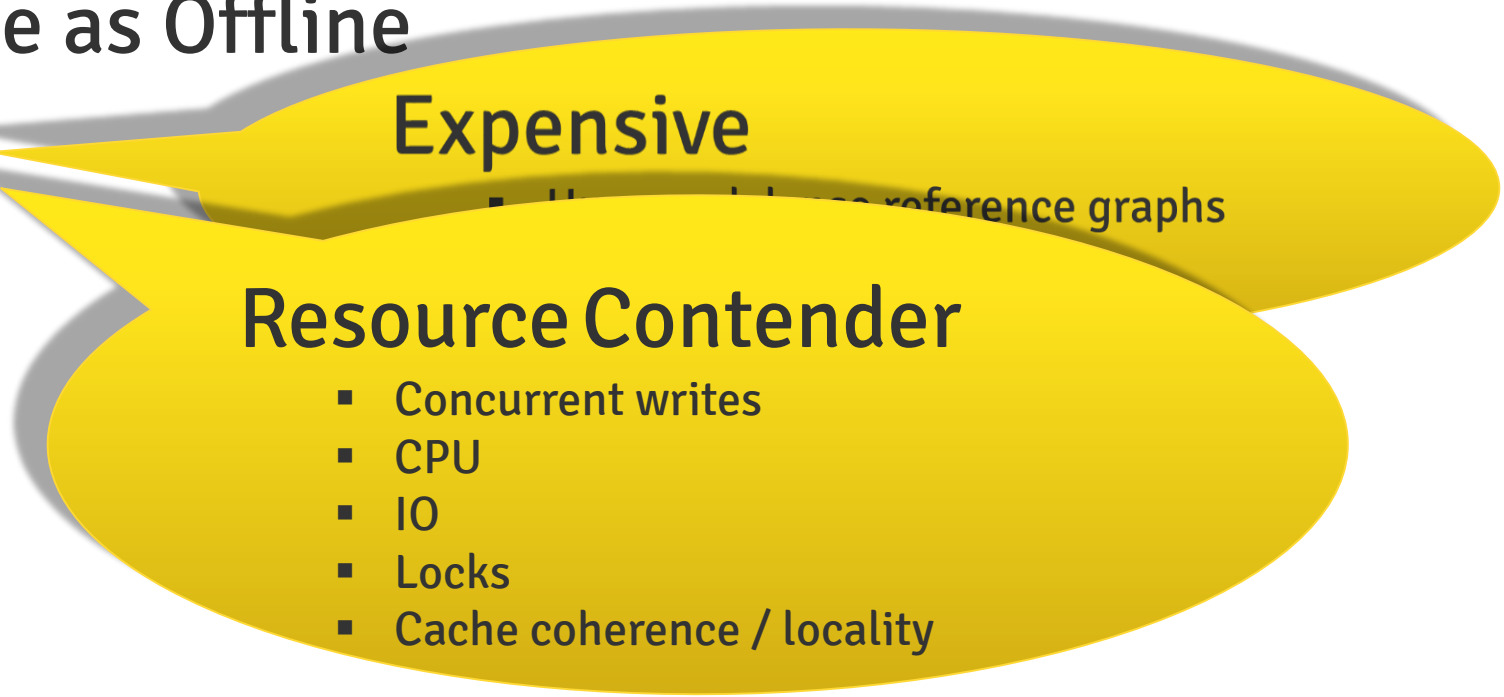
- Same as Offline
- BUT

Online Revisions GC

- Same as Offline
- BUT

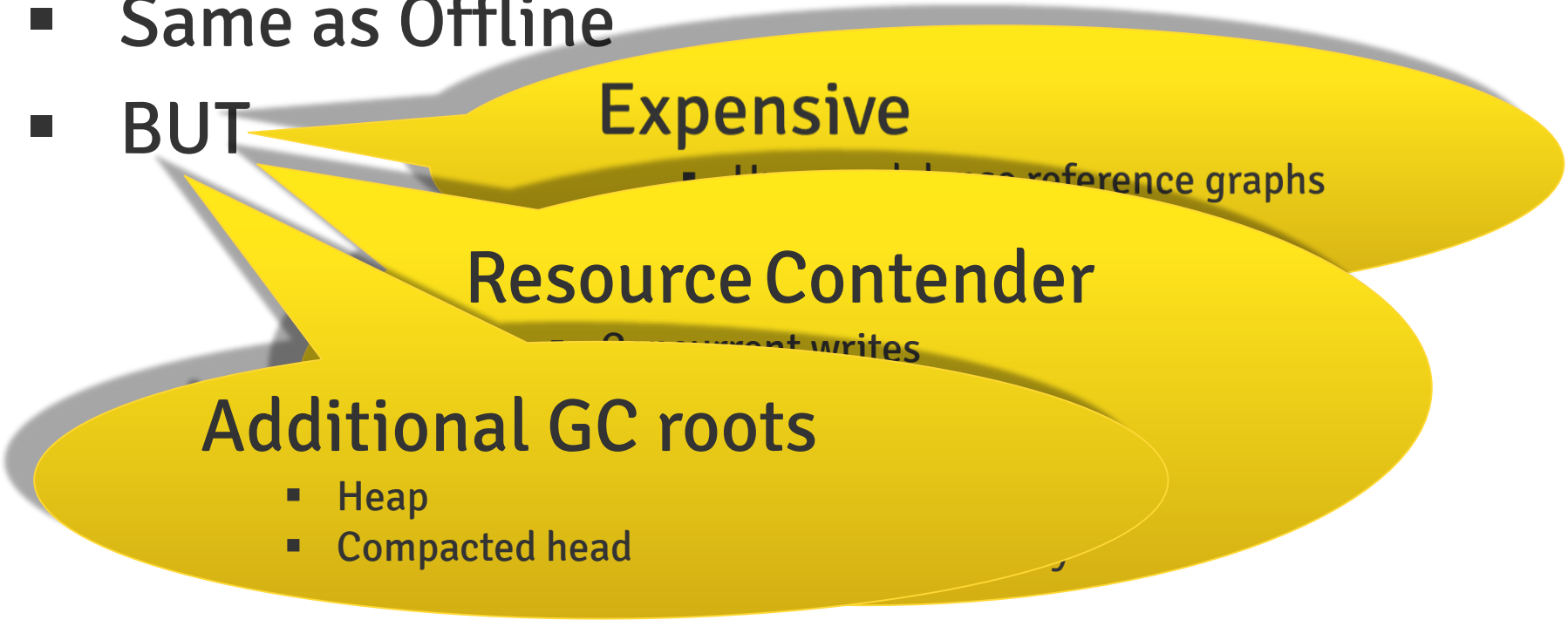
Online Revisions GC

- Same as Offline
- BUT

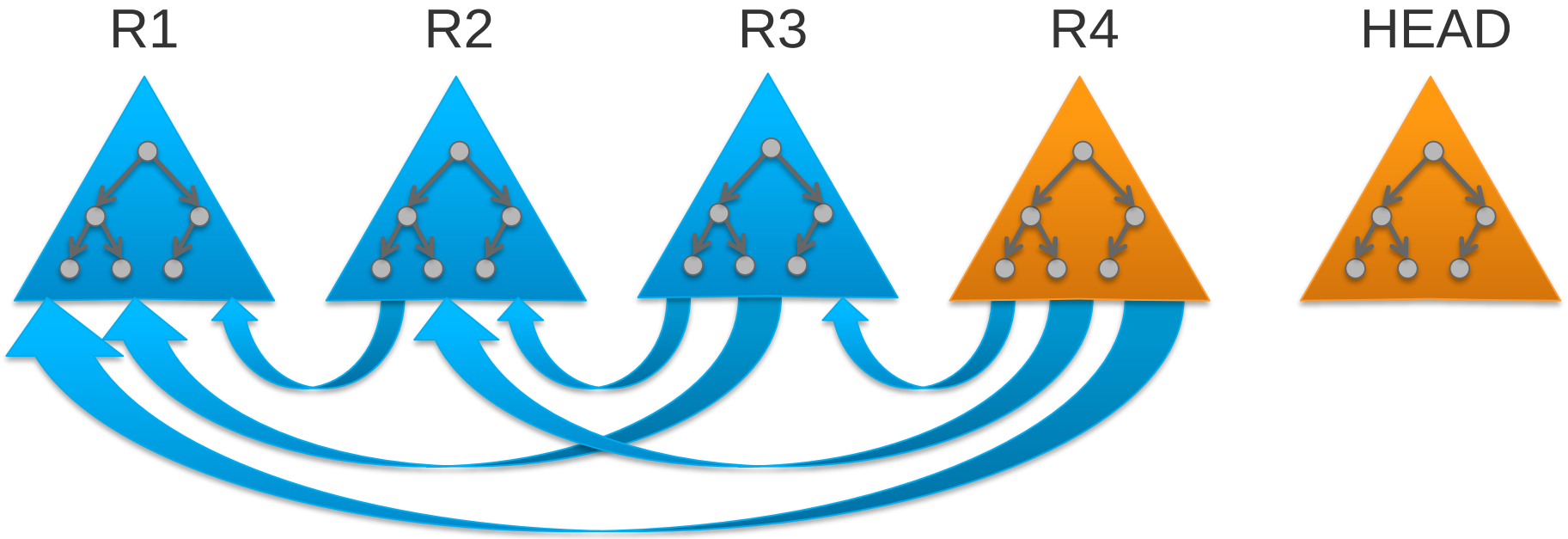


Online Revisions GC

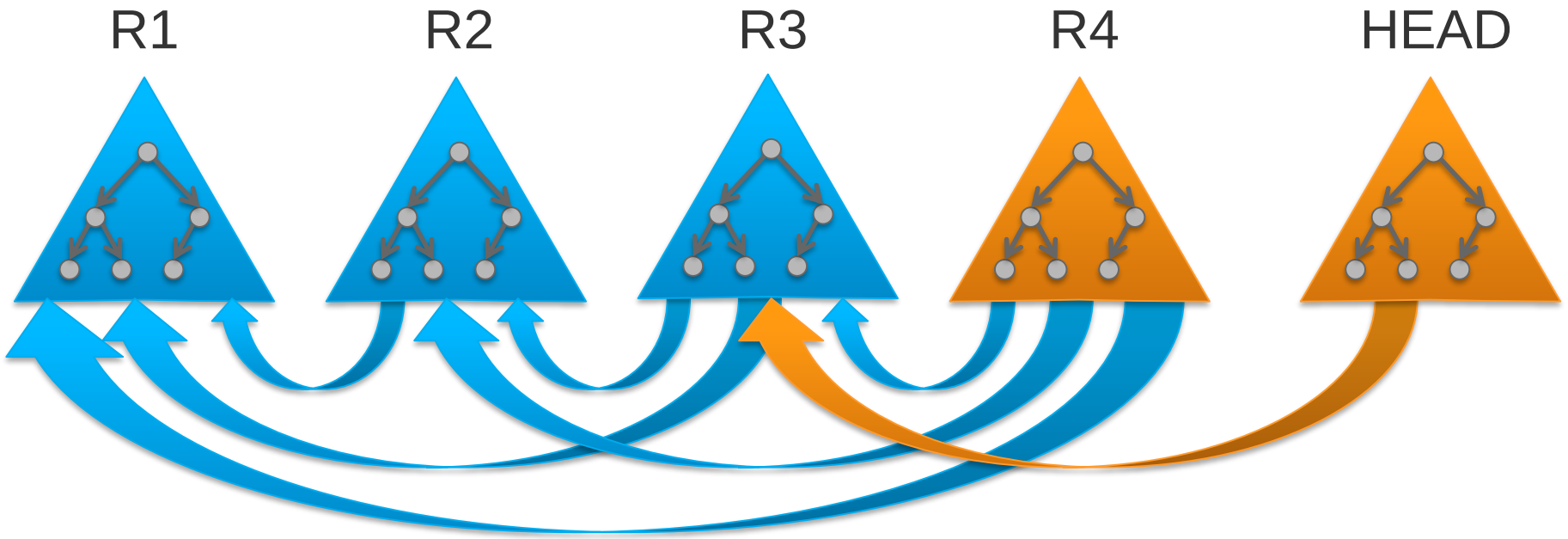
- Same as Offline
- BUT



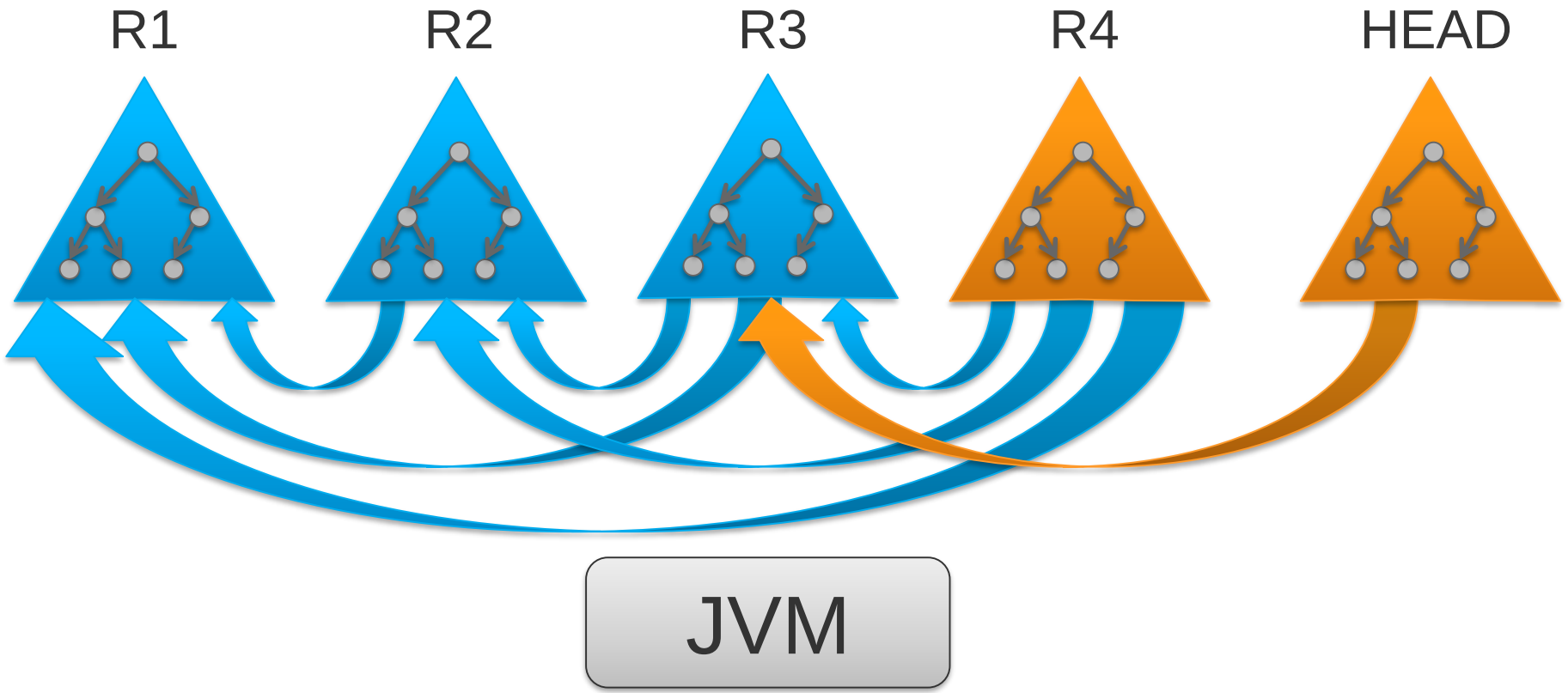
Online Revisions GC: Roots



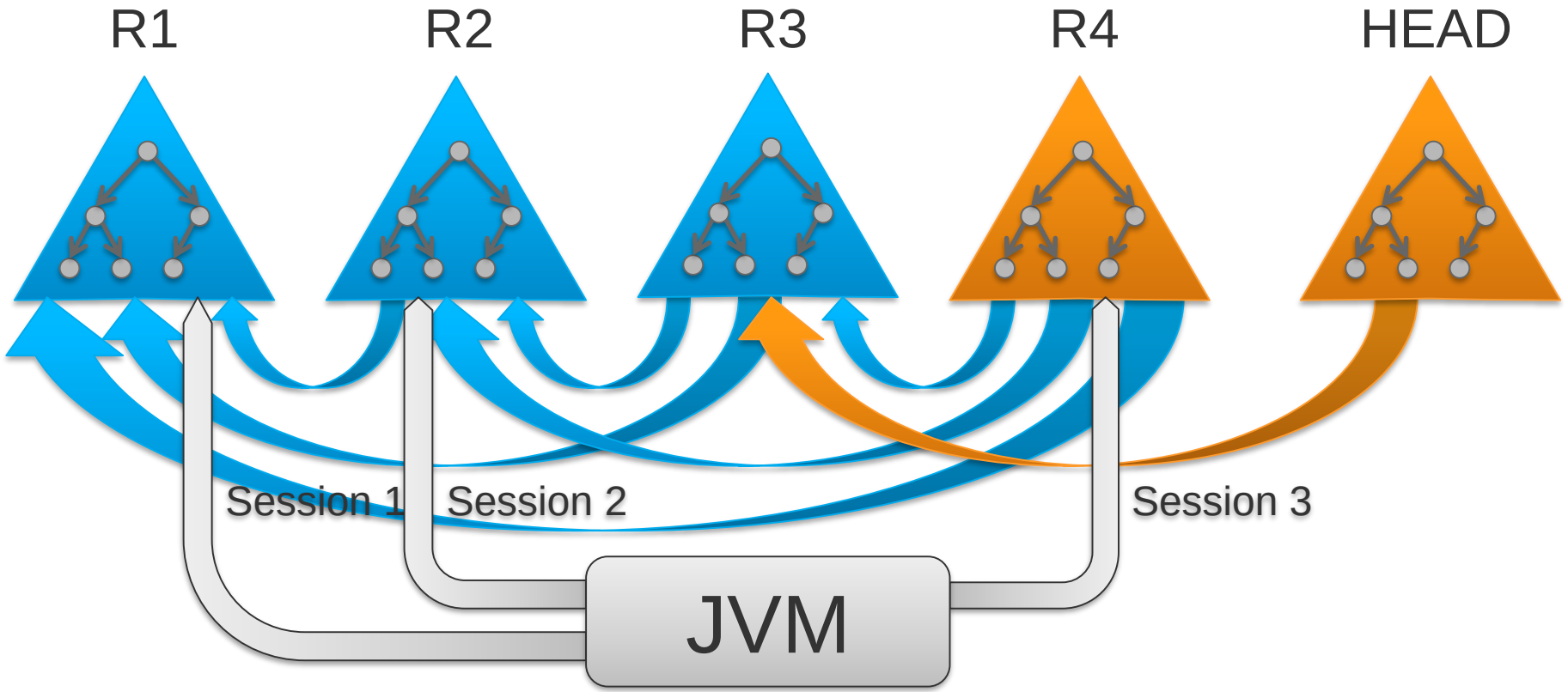
Online Revisions GC: Roots



Online Revisions GC: Roots



Online Revisions GC: Roots



Upcoming Improvements

Online Revisions GC

- Retention by Generation
  - No gc roots from compacted head
  - Ignore gc roots from heap
    - Configurable retention time instead
  - Cheaper than by reference

and beyond...

- Partial and background gc
  - Scalable
  - Resumable
  - Tunable

Conclusion

- Many improvements for gc
  - Changed storage format
  - Migration required
- Ground work for future improvements
  - More scalable gc
  - Bigger repositories
  - More write throughput

Questions