Map-Reduce with CouchDB

Kore Nordmann
<kore@php.net>
@koredn

March 21, 2010

Outline

Our documents

- Example: Issue tracker bug document

```
{ "_id":        "issue-42",
  "type":       "issue",
  "title":      "Bug in module foo",
  "text":       "Your software sucks!",
  "creator":    "user-bar",
  "state":      "new",
  "edited":     1935678239,
   ...
}
```

Our documents

- Example: Issue tracker bug comment

```
{ "_id":        "80572348asdfg789342",
  "type":       "issue-comment",
  "issue":      "issue-42",
  "edited":     1935678239,
  "text"   :    "I second that!",
   ...
}
```

Outline

Map-reduce-views

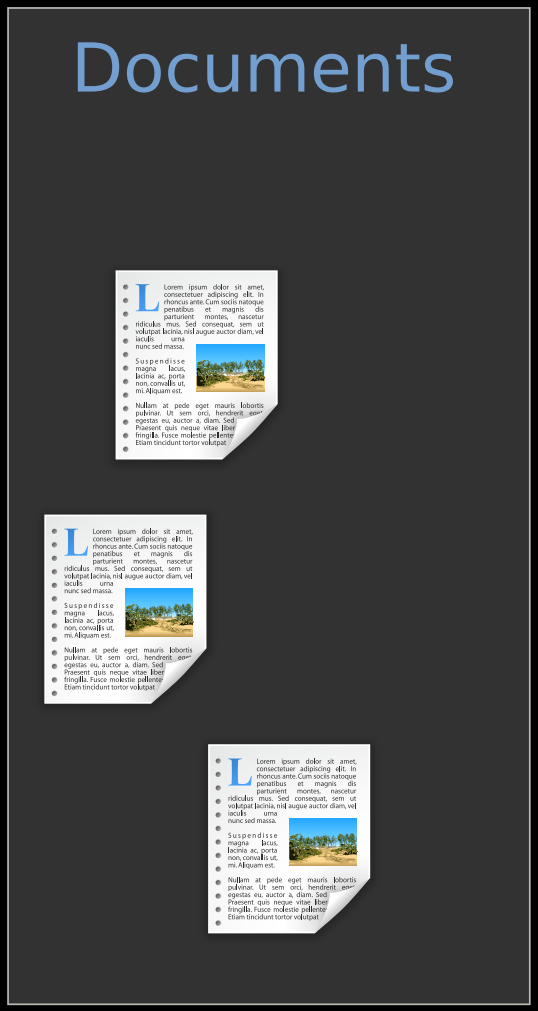
Map-reduce-views

- “MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.” [Wik09]
- Used by CouchDB to implement views

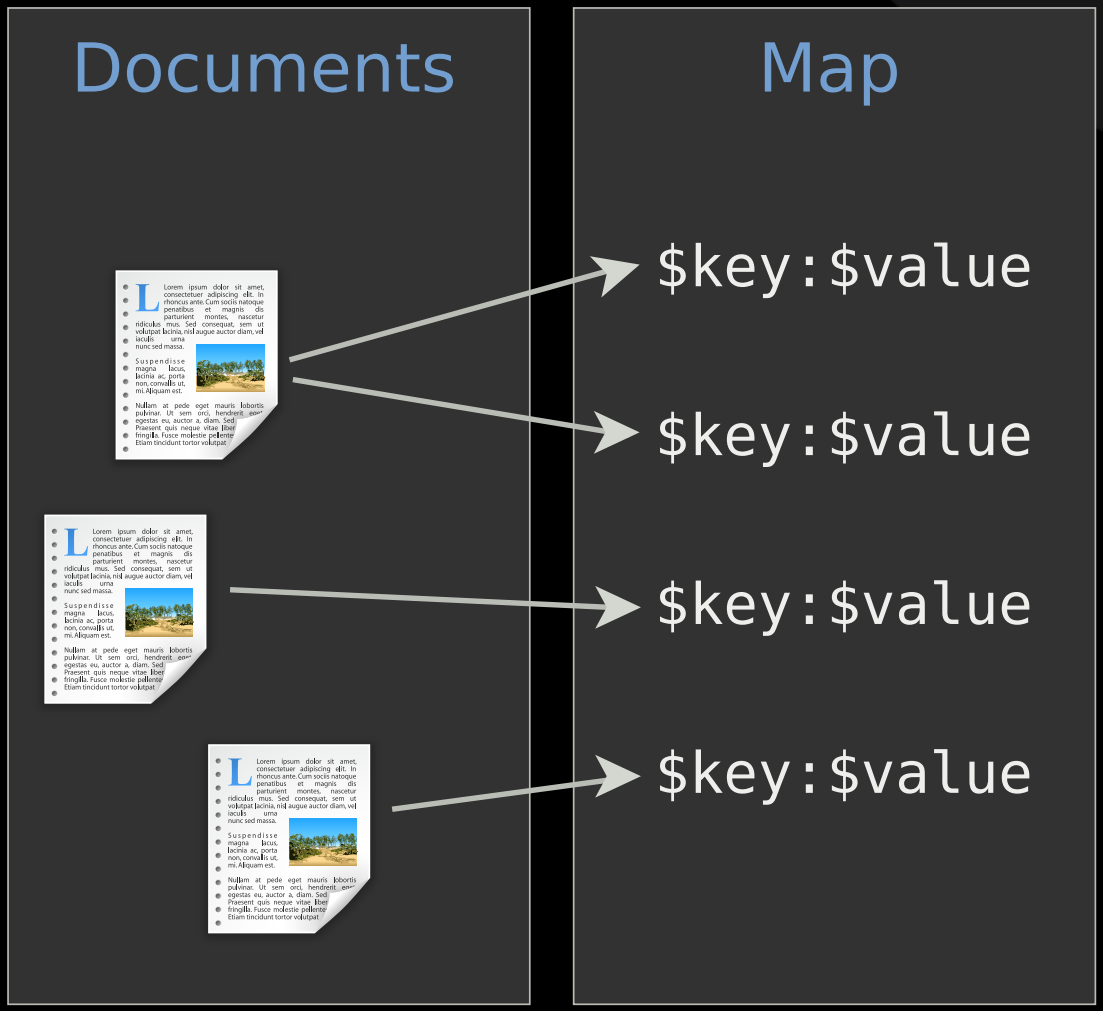
Map-reduce-views

- “MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.” [Wik09]
- Used by CouchDB to implement views
- Just a framework / pattern: You can implement “any” algorithm using map-reduce.

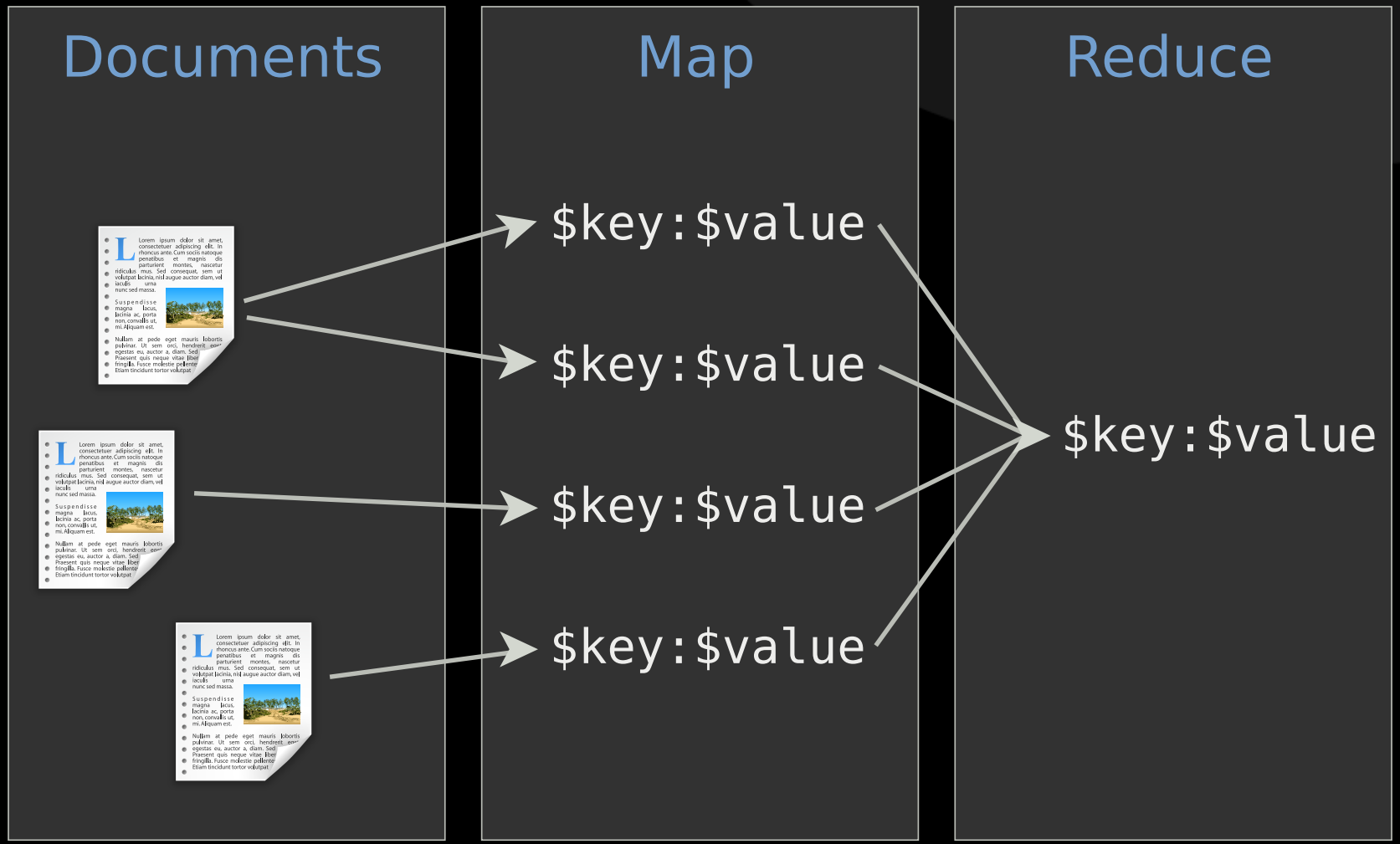
Map-Reduce



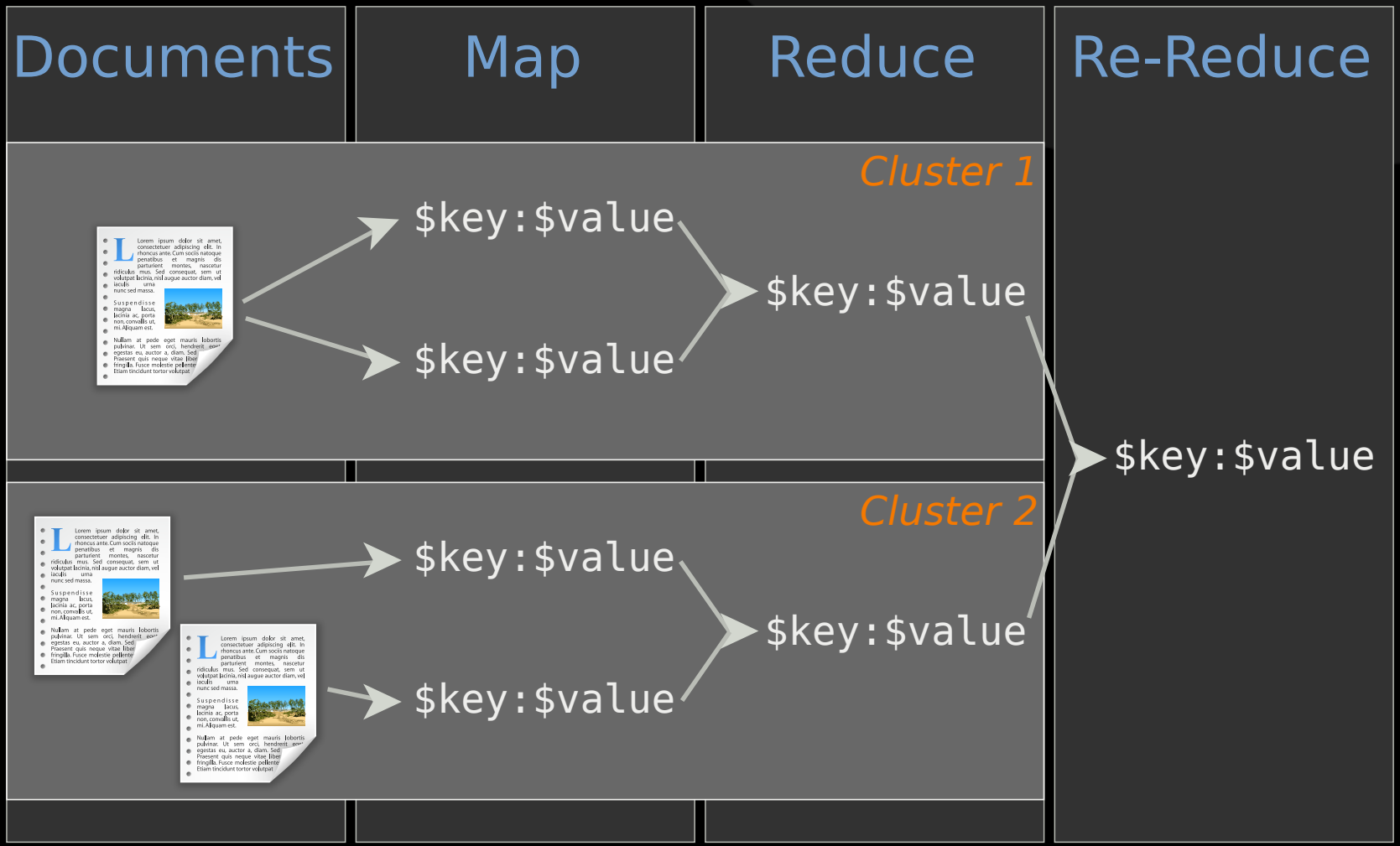
Map-Reduce



Map-Reduce


Documents
Map
Reduce
$key:$value
$key:$value
$key:$value
$key:$value
$key:$value

Map-Reduce


Documents
Map
Reduce
Re-Reduce
Cluster 1
$key:$value
$key:$value
$key:$value
Cluster 2
$key:$value
$key:$value
$key:$value
$key:$value

Map-Reduce

- Map and reduce functions are custom

Map-Reduce

- Map and reduce functions are custom
- Reduce is optional, plain view serves as a document index

Map-Reduce

- Map and reduce functions are custom
- Reduce is optional, plain view serves as a document index
- Reduce may be applied to subsets of the documents

Map-Reduce

- Map and reduce functions are custom
- Reduce is optional, plain view serves as a document index
- Reduce may be applied to subsets of the documents
- Reduce may be grouped

Our documents

- Example: Issue tracker bug document

```
{ "_id":        "issue-42",
  "type":       "issue",
  "title":      "Bug in module foo",
  "text":       "Your software sucks!",
  "creator":    "user-bar",
  "state":      "new",
  "edited":     1935678239,
   ...
}
```

Issue tracker

- Count issue states

```
1  function( doc )
2  {
3      if ( doc.type == "issue" )
4      {
5          emit( doc.state, 1 );
6      }
7  }
```

Map-Reduce example

- The simplest reduce function is just `count()`
  - Often used for statistics

```
1 function( keys, values, combine )
2 {
3     return sum( values );
4 }
```

Map-Reduce example

- The mapping result

```
1 "new"      => 1,
2 "assigned" => 1,
3 "new"      => 1,
4 "closed"   => 1,
5 "new"      => 1,
6 "new"      => 1,
7 "closed"   => 1,
8 "assigned" => 1,
9 "assigned" => 1,
```

Map-Reduce example

- The reduce result

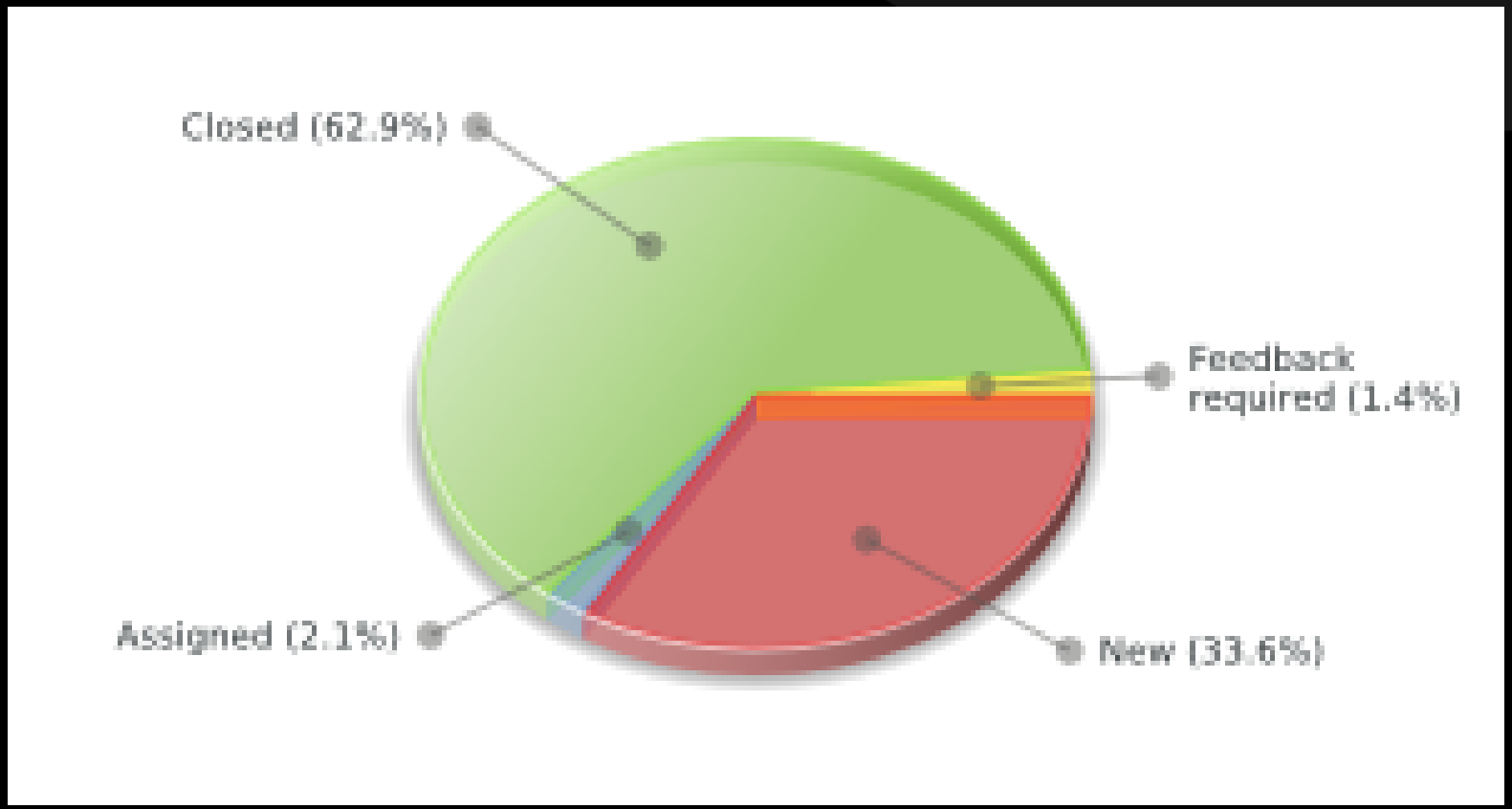
```
1  null          => 12
```

Map-Reduce example

- The grouped reduce result

```
1 "new"      => 5,
2 "assigned" => 23,
3 "closed"   => 42,
```

Result



Map-Reduce example

- The map function

```
1  function( doc )
2  {
3      if ( doc.type == "wiki" )
4      {
5          date = new Date();
6          date.setTime( doc.edited * 1000 );
7          emit( [
8                  date.getUTCFullYear(),
9                  date.getUTCMonth() + 1,
10                 date.getUTCDate(),
11                 date.getUTCHours(),
12                 date.getUTCMinutes(),
13                 date.getUTCSeconds(),
14             ], 1 );
15         // You could also emit the whole doc as value
16     }
17 }
```

Map-Reduce example

- The mapping result

```
[2008, 10, 11, 9,  11, 12] => 1
[2008, 10, 11, 9,  11, 12] => 1
[2008, 10, 11, 9,  11, 12] => 1
[2008, 10, 11, 9,  13, 8]  => 1
[2008, 10, 11, 9,  13, 44] => 1
[2008, 10, 11, 9,  14, 2]  => 1
[2008, 10, 12, 17, 46, 15] => 1
[2008, 10, 12, 17, 57, 52] => 1
[2008, 10, 12, 18, 0,  45] => 1
[2008, 10, 14, 8,  36, 29] => 1
[2008, 10, 14, 19, 33, 21] => 1
[2008, 10, 14, 19, 33, 35] => 1
```

Map-Reduce example

- The reduce function

```
1  function( keys, values, combine )
2  {
3      return sum( values );
4  }
```

Map-Reduce example

- The reduce result

```
1  null                                    => 12
```

Map-Reduce example

- The grouped reduce result

```
[2008, 10, 11, 9,  11, 12] => 3
[2008, 10, 11, 9,  13, 8]  => 1
[2008, 10, 11, 9,  13, 44] => 1
[2008, 10, 11, 9,  14, 2]  => 1
[2008, 10, 12, 17, 46, 15] => 1
[2008, 10, 12, 17, 57, 52] => 1
[2008, 10, 12, 18, 0,  45] => 1
[2008, 10, 14, 8,  36, 29] => 1
[2008, 10, 14, 19, 33, 21] => 1
[2008, 10, 14, 19, 33, 35] => 1
```

Map-Reduce example

- The filtered grouped reduce result
- `startkey=[2008,10,11]` and `endkey=[2008,10,12]`

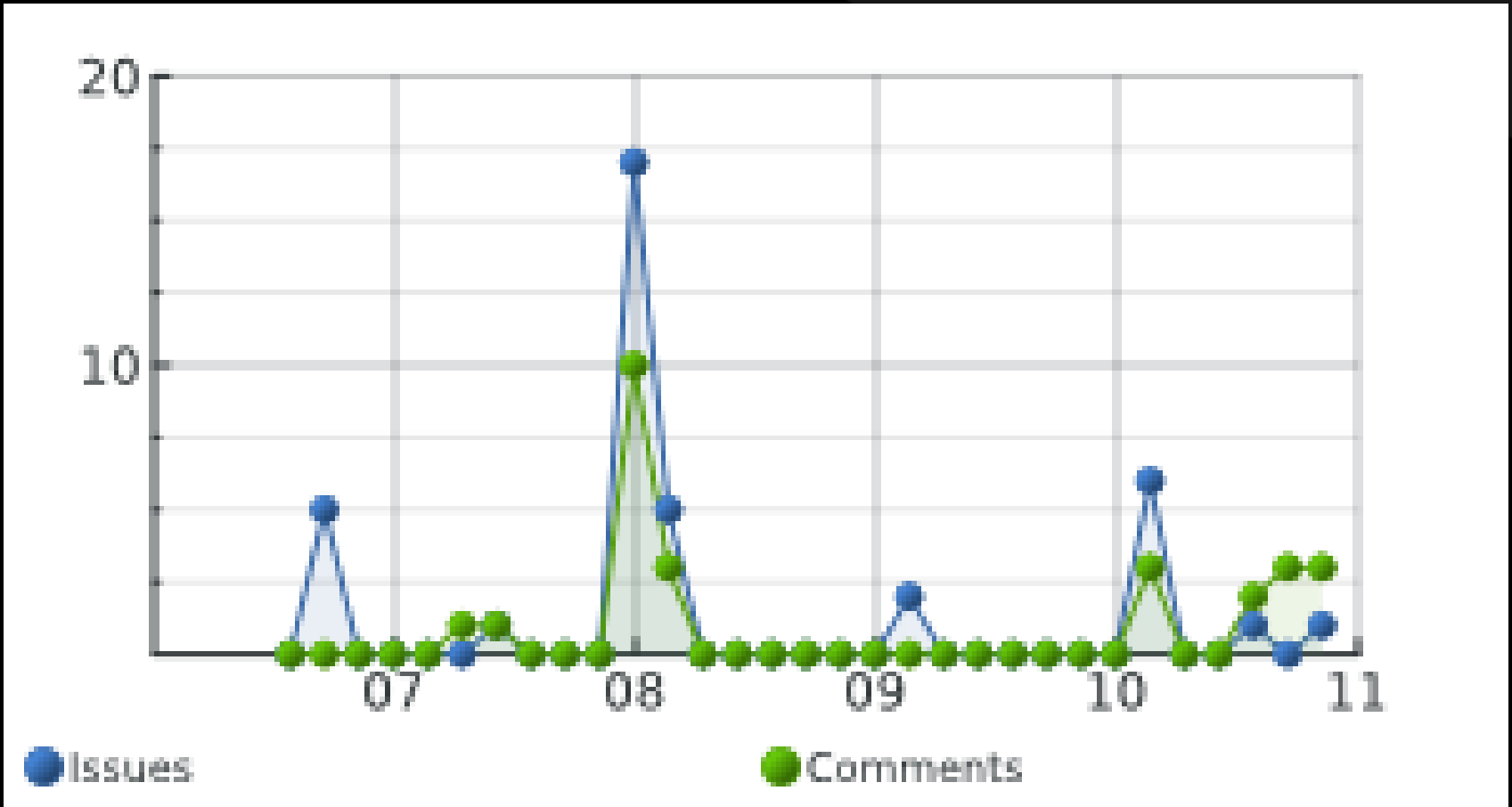
```
1 [2008, 10, 11, 9,  11, 12] => 3
2 [2008, 10, 11, 9,  13, 8]  => 1
3 [2008, 10, 11, 9,  13, 44] => 1
4 [2008, 10, 11, 9,  14, 2]  => 1
```

Map-Reduce example

- The grouped reduce result, with group level
- `group-level=3`

```
1 [2008, 10, 11]          => 6
2 [2008, 10, 12]          => 3
3 [2008, 10, 14]          => 3
```

Result



Full-Text-Search

- Index all documents by all their words

```
1  function( doc ) {
2    if ( doc.type == "issue" ) {
3      // Simple word indexing, does not respect overall
           occurences of words,
4      // stopwords, different word seperation characters,
           or word variations.
5      var text  = doc.title.replace( /[\s:.,!?-]+/g, "␣"
           ) +
6                  doc.text.replace( /[\s:.,!?-]+/g, "␣" )
                       ;
7      var words = text.split( "␣" );
8      for ( var i = 0; i < words.length; ++i ) {
9        value = {};
10       value[doc._id] = 1;
11       emit( words[i].toLowerCase(), value );
12     }
13   }
14 }
```

Issue tracker

- Index all documents by all their words

```
1  ...
2  "a"     => {issue-8: 1}
3  "a"     => {issue-8: 1}
4  "a"     => {issue-8: 1}
5  "a"     => {issue-8: 1}
6  "a"     => {issue-81: 1}
7  "a"     => {issue-83: 1}
8  "a"     => {issue-83: 1}
9  "able"  => {issue-39: 1}
10 "able"  => {issue-56: 1}
11 "able"  => {issue-73: 1}
12 "able"  => {issue-80: 1}
13 "about" => {issue-24: 1}
14 "about" => {issue-43: 1}
15 "about" => {issue-85: 1}
16 ...
```

Full-Text-Search

- Reduce by word count

```
1  function( keys, values ) {
2    var count = {};
3    for ( var i in values ) {
4      for ( var id in values[i] ) {
5        if ( count[id] ) {
6          count[id] = values[i][id] + count[id];
7        } else {
8          count[id] = values[i][id];
9        }
10     }
11   }
12   return count;
13 }
```

Issue tracker

- Index all documents by all their words

```
...
"a"       => {
                issue-68: 6,
                issue-66: 6,
                issue-22: 4,
                issue-63: 3,
                issue-60: 2,
                issue-35: 2,
                issue-34: 1,
                issue-31: 1,
                ...
              }
"able"    => {issue-86: 1, issue-80: 1, issue-73: 1,
    issue-56: 1, issue-39: 1}
"about"   => {issue-85: 1, issue-43: 1, issue-24: 1}
...
```

Outline

Inter document links

- There is no ensured inter document consistency in CouchDB

ents:

Inter document links

- There is no ensured inter document consistency in CouchDB
- Different possibilities of relating documents:

Inter document links

- There is no ensured inter document consistency in CouchDB
- Different possibilities of relating documents:
  - List IDs of related documents in document (n:m)

Inter document links

- There is no ensured inter document consistency in CouchDB
- Different possibilities of relating documents:
  - List IDs of related documents in document (n:m)
  - ... both directions are feasible

Inter document links

- There is no ensured inter document consistency in CouchDB
- Different possibilities of relating documents:
  - List IDs of related documents in document (n:m)
  - ... both directions are feasible
  - Embed the whole related document (1:n)

Inter document links

- There is no ensured inter document consistency in CouchDB
- Different possibilities of relating documents:
  - List IDs of related documents in document (n:m)
  - ... both directions are feasible
  - Embed the whole related document (1:n)
- Solution depends on update-ratio

```
{ "type":  "issue",
  "title": "Hello world",
  "text":  "...",
  "comments": [
    { "comment": "..." },
  ],
  "creator": "user-foo",
}
```

Issue tracker

- JOIN query

```
1  function( doc )
2  {
3      if ( doc.type == "issue" )
4      {
5          emit( [doc._id, 0], doc._id );
6      }
7
8      if ( doc.type == "issue_comment" )
9      {
10         emit( [doc.issue, doc._id], doc._id );
11     }
12 }
```

Issue tracker

- JOIN query result

```
1 ["tracker_issue-1", 0] => "tracker_issue-1"
2 ["tracker_issue-1", "1eaaa503adbec07b0013c060b5d7b53c"]
      => "1eaaa503adbec07b0013c060b5d7b53c"
3 ["tracker_issue-1", "6e77ebd5ea0383cf1d8ed505f2517ca7"]
      => "6e77ebd5ea0383cf1d8ed505f2517ca7"
4 ["tracker_issue-1", "b0502afb84a80a48fecc7442153c0aa2"]
      => "b0502afb84a80a48fecc7442153c0aa2"
5 ["tracker_issue-10", 0] => "tracker_issue-10"
6 ["tracker_issue-10", "63872cd8a4d5301238dbdb084a4d7a3f"
    ] => "63872cd8a4d5301238dbdb084a4d7a3f"
7 ["tracker_issue-10", "9af0dd3c4d184c07c2ff98982f98d39f"
    ] => "9af0dd3c4d184c07c2ff98982f98d39f"
```

- Can again be filtered...
- Using "`?include_docs=true`" also gives you all documents (since 0.11)

Outline

Inter document links

```
{ "type":     "forum-post",
  "title":    "Hello world",
  "text":     "...",
  "parents": [ "forum-post-23", "forum-post-42" ]
  "creator": "user-foo",
}
```

Issue tracker

- Tree view

```
1  function( doc )
2  {
3      if ( doc.type == "forum-post" )
4      {
5          location = doc.parents;
6          location.append( doc._id );
7          emit( location, doc._id );
8      }
9  }
```

Issue tracker

- Tree query result

```
1  ["forum-post-1"]
2      => "forum-post-1"
3  ["forum-post-1", "forum-post-2"]
4      => "forum-post-2"
5  ["forum-post-1", "forum-post-2", "forum-post-4"]
6      => "forum-post-4"
7  ["forum-post-1", "forum-post-3"]
8      => "forum-post-3"
```

Issue tracker

- Tree query subtree result
- `startkey=["forum-post-1", "forum-post-2"]`

```
1 ["forum-post-1", "forum-post-2"]
2     => "forum-post-2"
3 ["forum-post-1", "forum-post-2", "forum-post-4"]
4     => "forum-post-4"
```

Outline

The end

- Open questions?
- Further remarks?
- Contact
  - Mail: <kore@php.net>
  - Web: `http://kore-nordmann.de/` (Slides will be available here soonish)
  - Twitter: `http://twitter.com/koredn`

Bibliography I

[Wik09] Wikipedia, *Mapreduce — wikipedia, the free encyclopedia*, 2009, [Online; accessed 27-August-2009].