Class 7: Pair RDDs and Joins

New York University

Summer 2017

- **Another example using `join`**

userID *requestedFile (contains DocID)*

webLogsFile

```
56.38.234.188 - 99788 "GET /KBDOC-00157.html HTTP/1.0" …
56.38.234.188 - 99788 "GET /theme.css HTTP/1.0" …
203.146.17.59 - 25254 "GET /KBDOC-00230.html HTTP/1.0" …
221.78.60.155 - 45402 "GET /titanic_4000_sales.html HTTP/1.0" …
65.187.255.81 - 14242 "GET /KBDOC-00107.html HTTP/1.0" …
…
```

- *Use `join` on `docID` to combine data*
  - *`docID` is embedded in the `requestedFile` field of the `webLogs` file*
  - *`docID` is its own field in the `kbList` file*

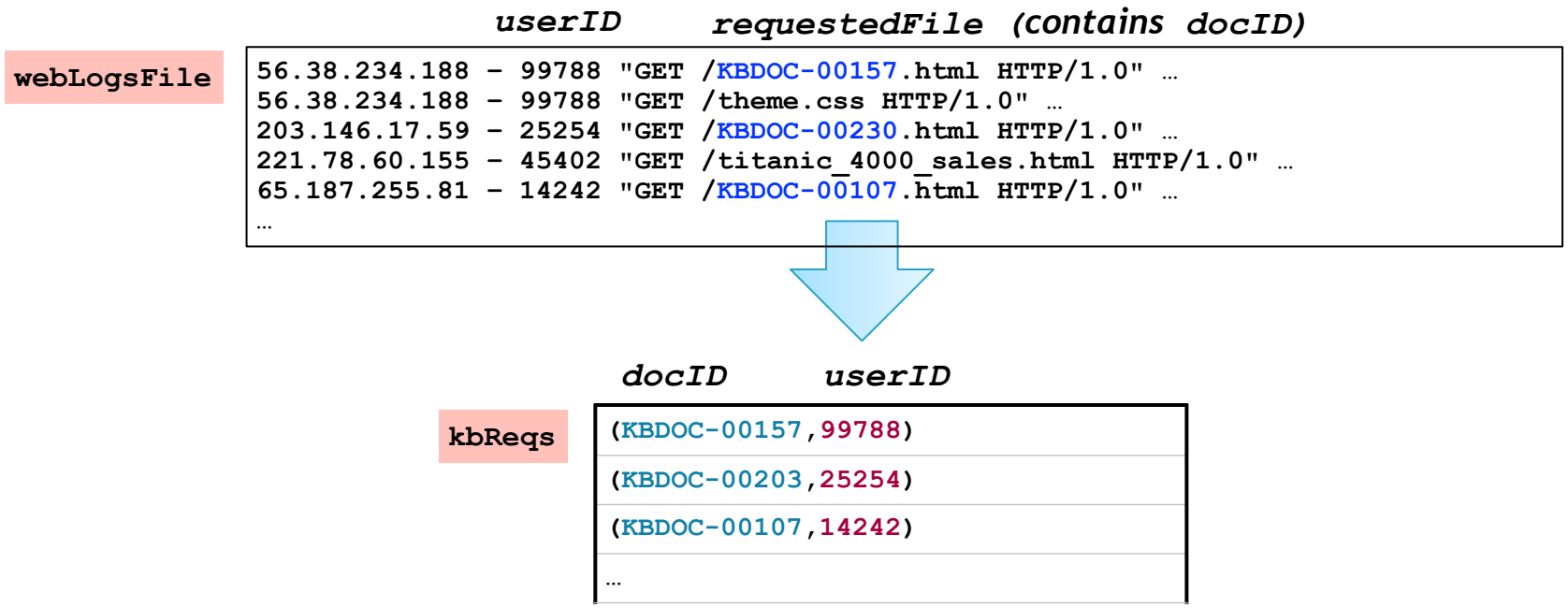
docID *Title*

kbList

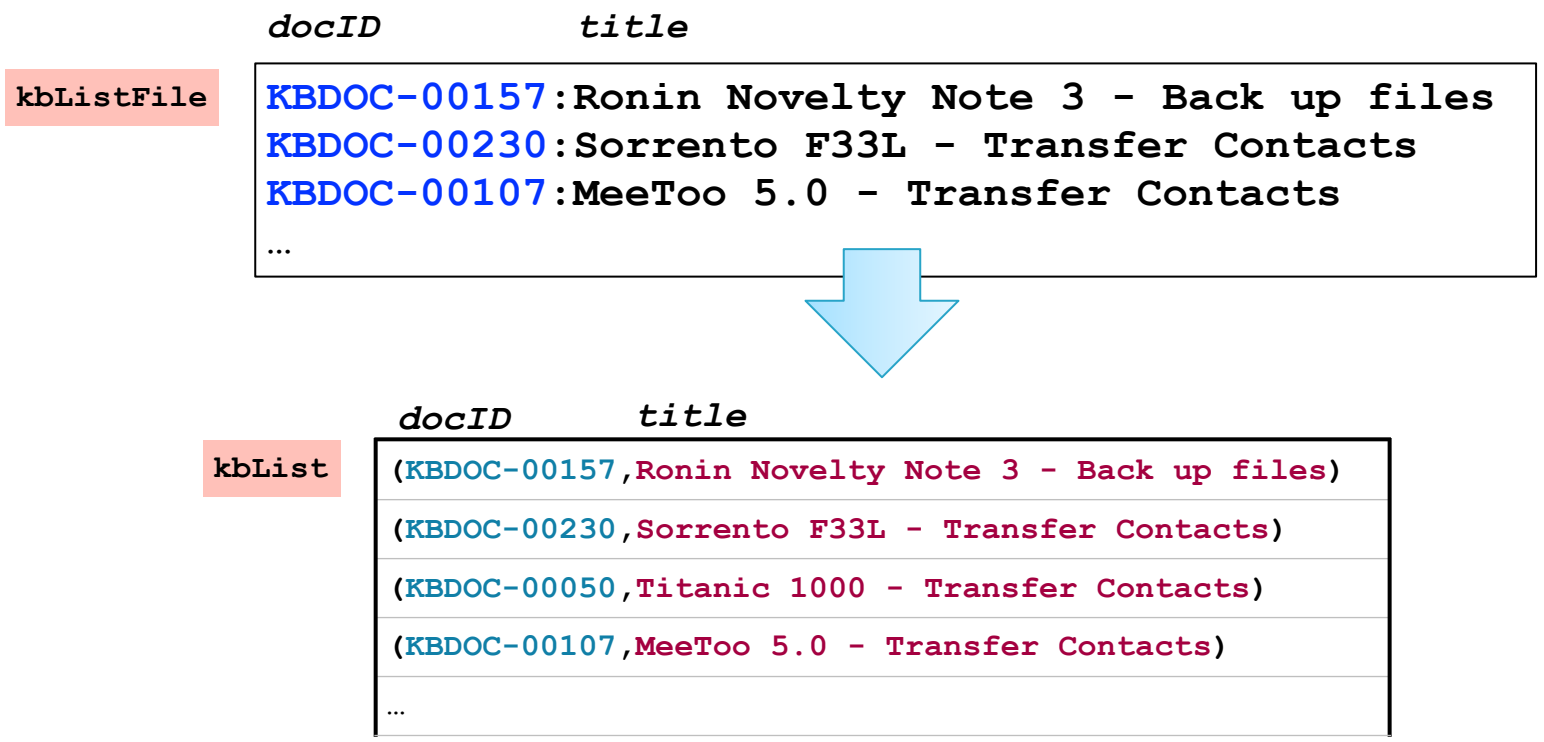
```
KBDOC-00157:Ronin Novelty Note 3 - Back up files
KBDOC-00230:Sorrento F33L - Transfer Contacts
KBDOC-00107:MeeToo 5.0 - Transfer Contacts
…
```

- **Step by step refinement**
  1. Map each dataset into key-value Pair RDDs
     a. Map requests in **`webLogs`** to **`(docID,userID)`**
     b. Map **`kbList`** to **`(docID,title)`**
  2. Join the two Pair RDDs on **`docID`**
  3. Map resulting RDD into final format: **`(userID,title)`**
  4. Group **`title`** by **`userID`**

- Load weblogs file into RDD **`webLogsFile`**
- Get only the lines that reference a knowledge base document - use **`filter`** to select only lines containing **`"KBDOC-"`**
- Get **`userID`**: Use **`map`** and **`split`** to split the line on ' ' and take the second element, which is **`userID`**
- Produce the RDD **`kbReqs`** that contains the above fields



- Load the knowledge base file into RDD **kbListFile**
- Separate **docID** and **title** into fields - use **map** and **split** to split on ':' and again use **map** to map the two fields to a pair of fields
- Store the result to RDD **kbList**



- Create RDD **titleReqs** using **join** on the **docID** field of RDDs **kbReqs** and **kbList**
- Use **map** to carry **userID** and **title** into the RDD **titleReqs**

kbReqs

docID	*userID*
(KBDOC-00157,	99788)
(KBDOC-00230,	25254)
(KBDOC-00107,	14242)
…	

kbList

docID	*title*
(KBDOC-00157,	Ronin Novelty Note 3 - Back up files)
(KBDOC-00230,	Sorrento F33L - Transfer Contacts)
(KBDOC-00050,	Titanic 1000 - Transfer Contacts)
(KBDOC-00107,	MeeToo 5.0 - Transfer Contacts)
…	

titleReqa

userID	*title*
(99788,	Ronin Novelty Note 3 - Back up files)
(25254,	Sorrento F33L - Transfer Contacts)
(14242,	MeeToo 5.0 - Transfer Contacts)
…	

- Finally, use **`groupByKey`** to group by **`userID`** value

userID	*title*
(99788,	Ronin Novelty Note 3 - Back up files)
(25254,	Sorrento F33L - Transfer Contacts)
(14242,	MeeToo 5.0 - Transfer Contacts)
…	

`titleReqa`

userID	*title*
(99788,	[Ronin Novelty Note 3 - Back up files, Ronin S3 - overheating])
(25254,	[Sorrento F33L - Transfer Contacts])
(14242,	[MeeToo 5.0 - Transfer Contacts, MeeToo 5.1 - Back up files, iFruit 1 - Back up files, MeeToo 3.1 - Transfer Contacts])
…	

Note: Values are grouped into Iterables.

Homework

See the homework packet for details.