junosearch Architecture Diagrams

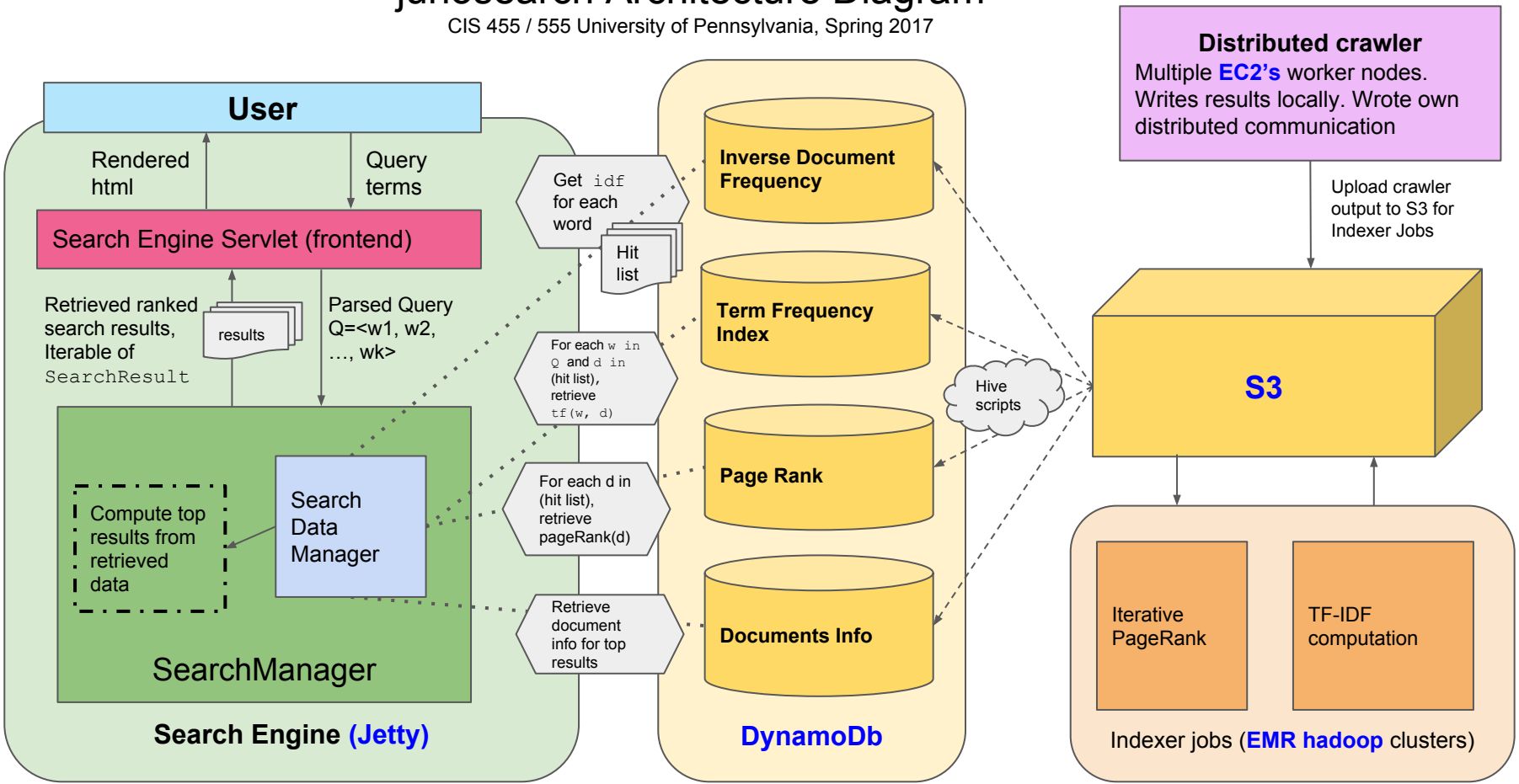
University of Pennsylvania, CIS, School of Engineering and Applied Sciences
CIS 455 / 555 Final Project

Wu Wai, Akash Subramanian, Victoria Xiao, Alex Zhao

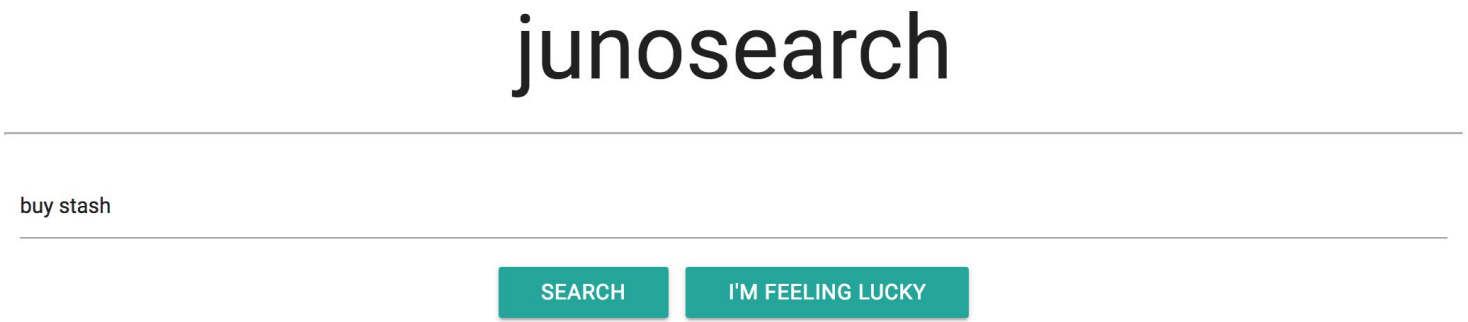
I. Overall Architecture

junosearch Architecture Diagram

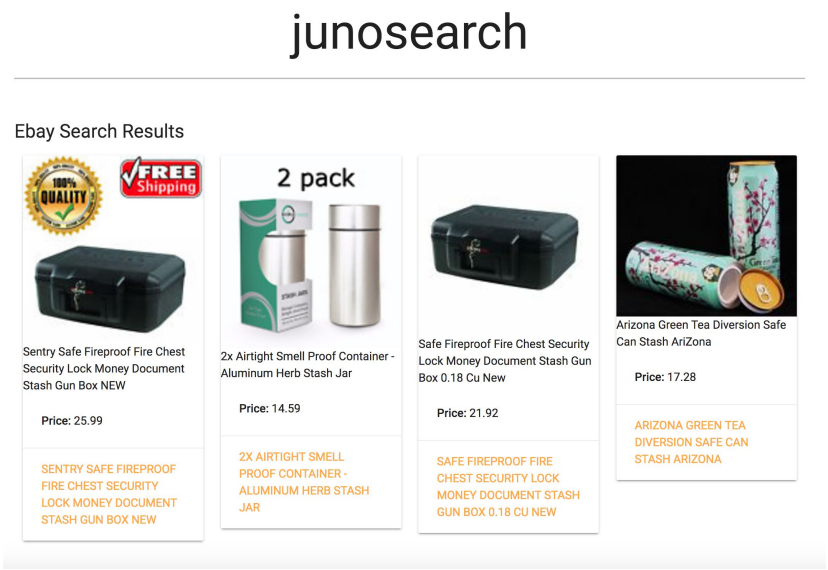
CIS 455 / 555 University of Pennsylvania, Spring 2017

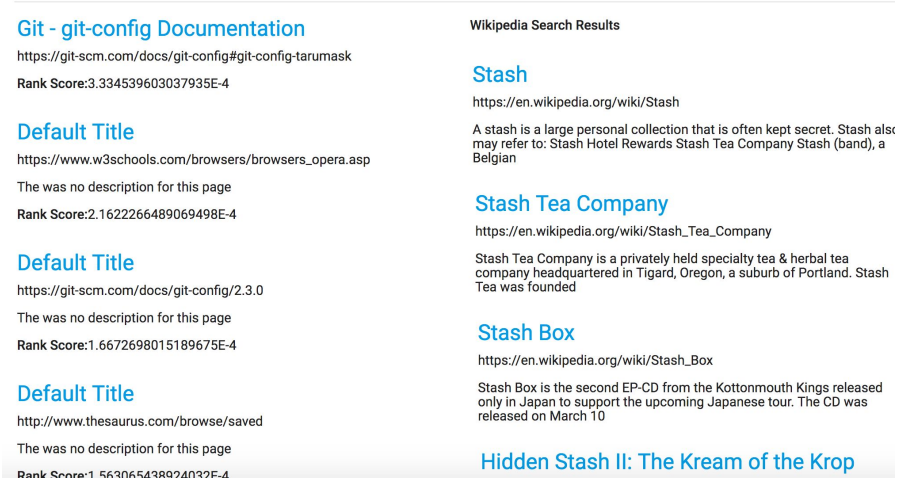


Front End Samples Demo (search page)



Front End Samples Demo (results page)





Database (DynamoDb) scheme

Documents_Info `(stores meta info about documents)`

```
Document {
        "url": "String"; // primary partition key
        "description": "String"; // document description
        "Title": "String"; // title of the document
}
```

Inverse_Document_Frequency `(idf table)`

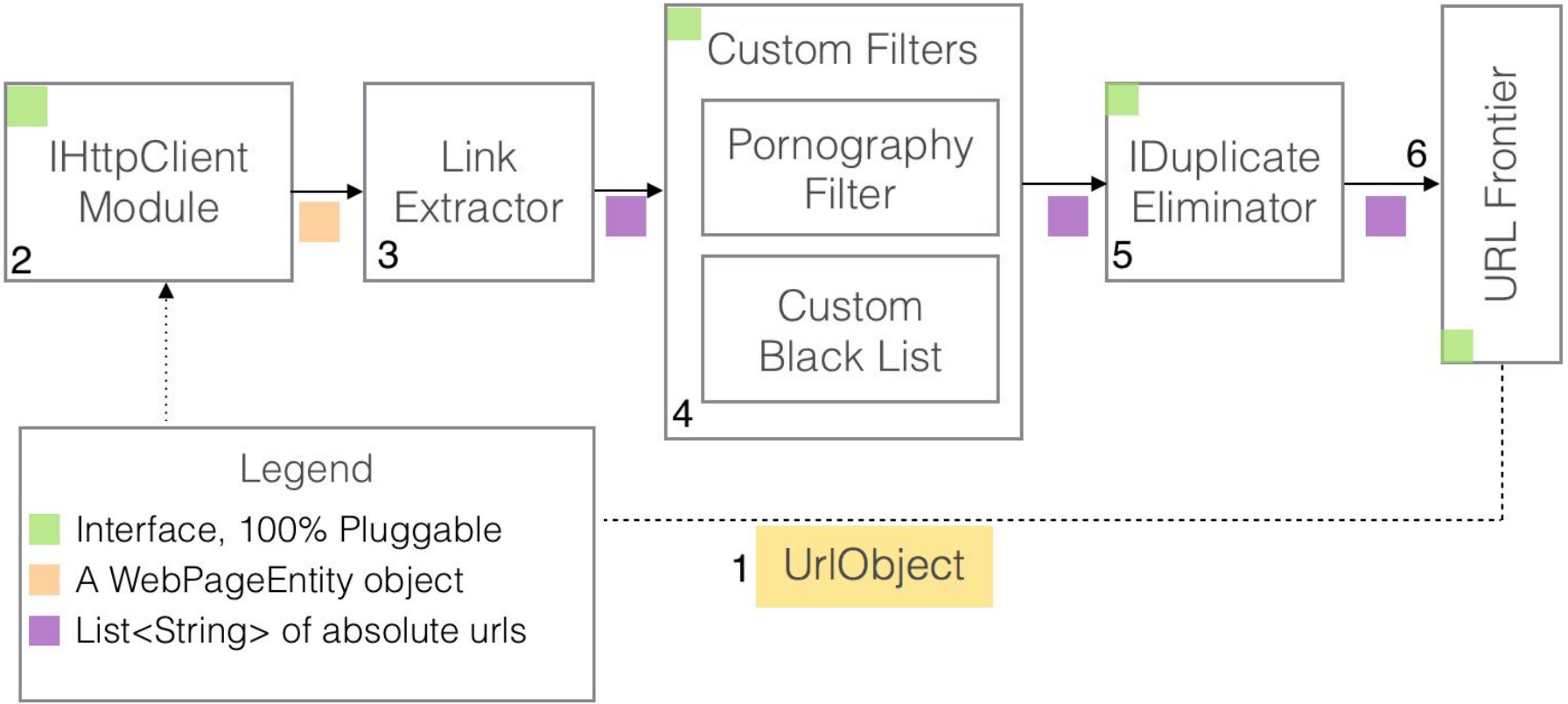
```
Document_Freqency {
        "term": "String"; // primary sort key
        "idf": "Number"; // primary partition key
}
```

Term_Frequency_Index `(stores tf (w, d) for each w,d pair)`

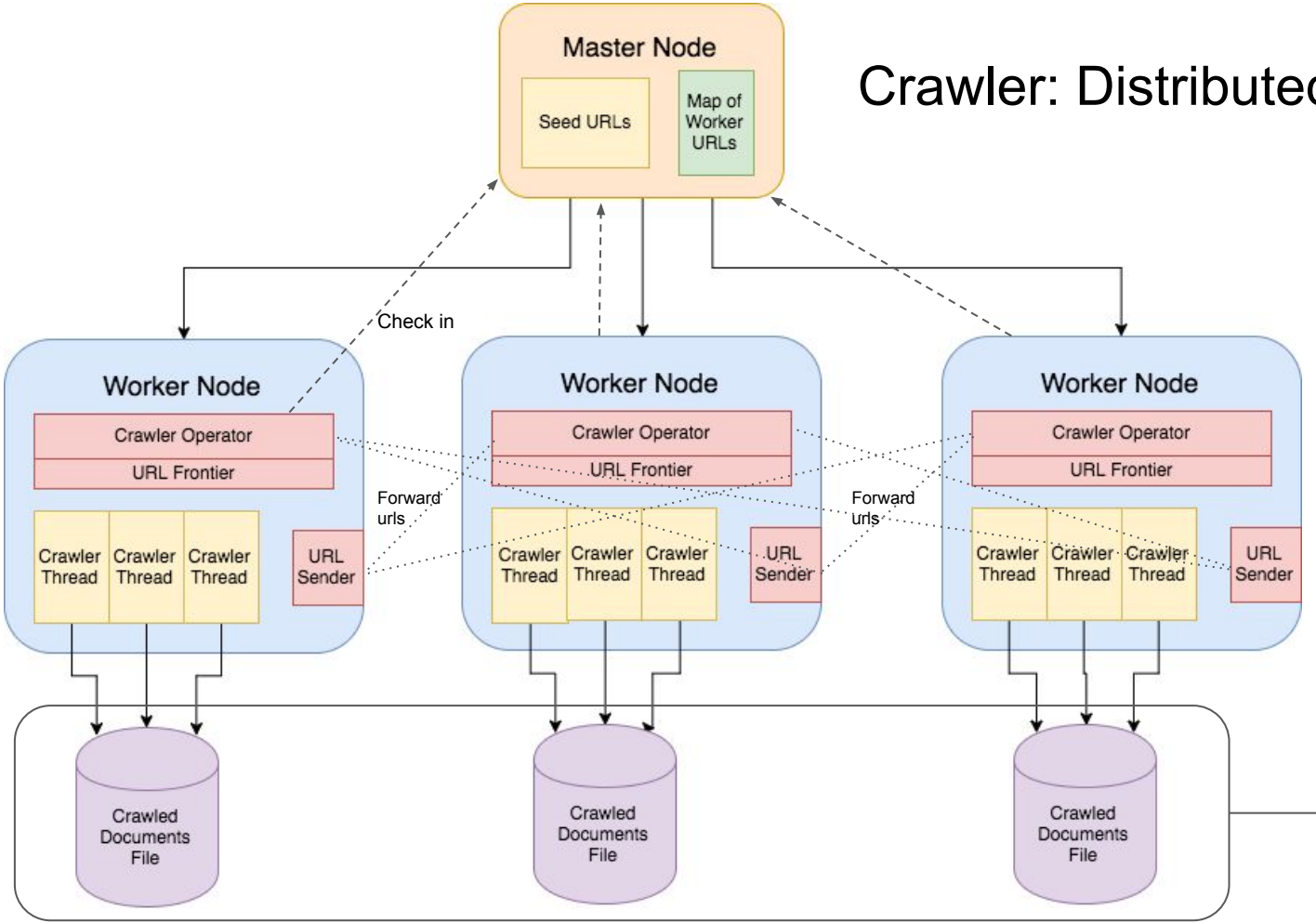
```
Term {
        "term": "String"; // primary partition key
        "tf": "Number"; // primary sort key

        // url of the document for which this tf applies
        "url": "String";

        // a list of where the term appears in the document
specified by the url. Originally for ranking purposes
        "positions": "List<Number>";

}
```

II. Crawler Diagrams

Crawler: Single Threaded, Single Node Design

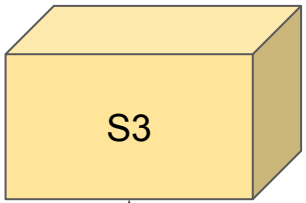


Crawler: Distributed Architecture



Notes: only 3 nodes shown out of 9 for simplicity. Number of threads per node anywhere from 10 to 500 during testing

Per thread operations similar to single threaded architecture



III. Indexer

MapReduce Job I - Computing Term Frequency

Mapper:

Key
line number

Value
<url>\t<content>\t<outlink1,2,...>\t<metadata>

Emit
(<word>\t<url>,
<term_frequency>-<position1,position2,...>)

Reducer:

Key
<term>\t<url>

Values [<term>-<position1,position2,...>]

Emit
(<term>\t<url>,
<term_frequency>\t<position1,position2,...>)

MapReduce Job II - Computing Document Frequency

Mapper:

Key
line number

Value
<term>\t<url>\t<tf>\t<position1,position2,...>

Emit
(<term>, “1”)

Reducer:

Key
<term>

Values
[“1”,..., “1”]

Emit
(<term>, <values_length>)

MapReduce Job III - Document Title and Description

Mapper:

Key
line number

Value
<url>\t<content>\t<outlink1,2,...>\t<metadata>

Emit
(<url>, <title>\t<description>)

Reducer:

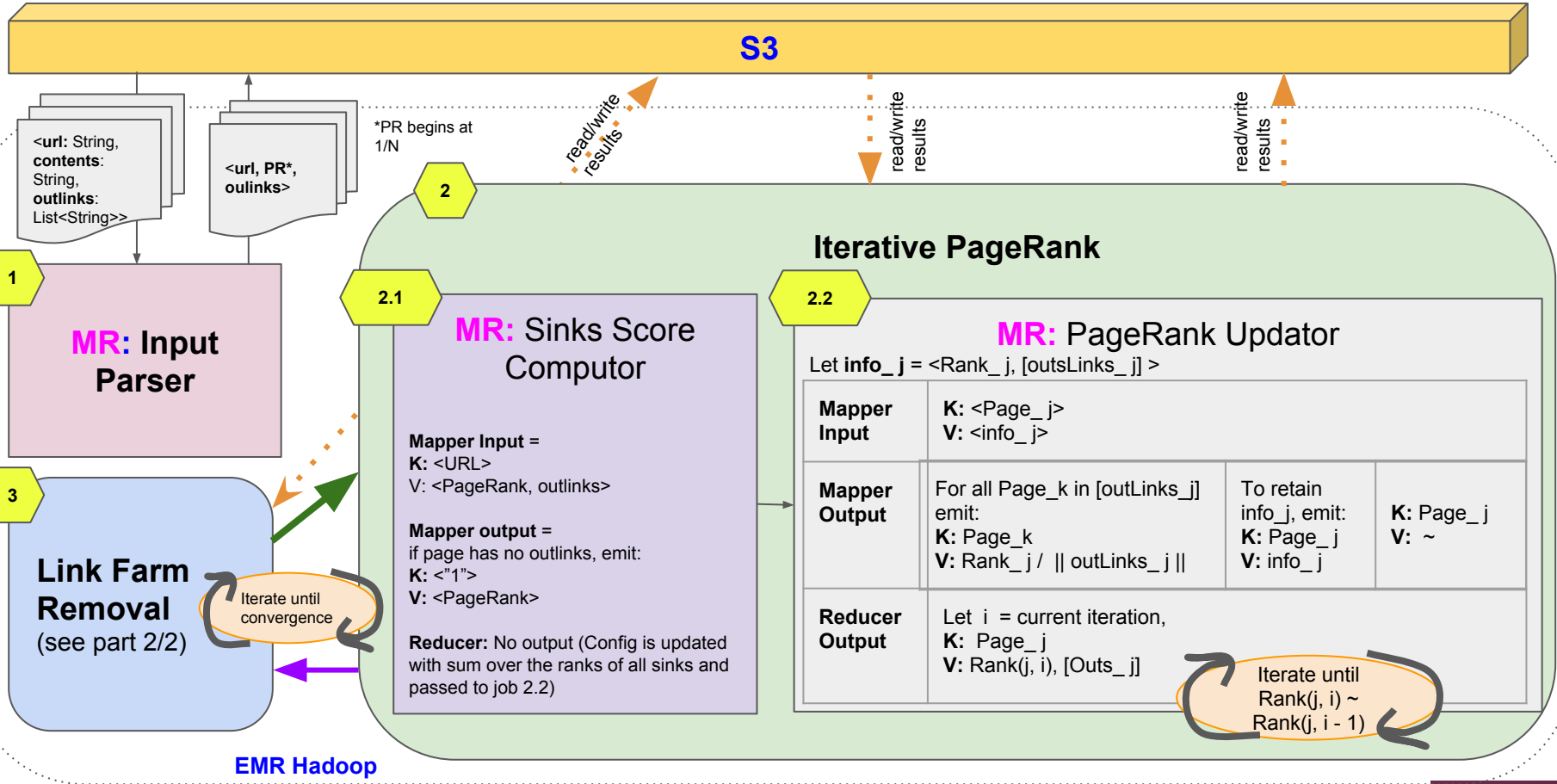
Key
<url>

Value <title>\t<description>

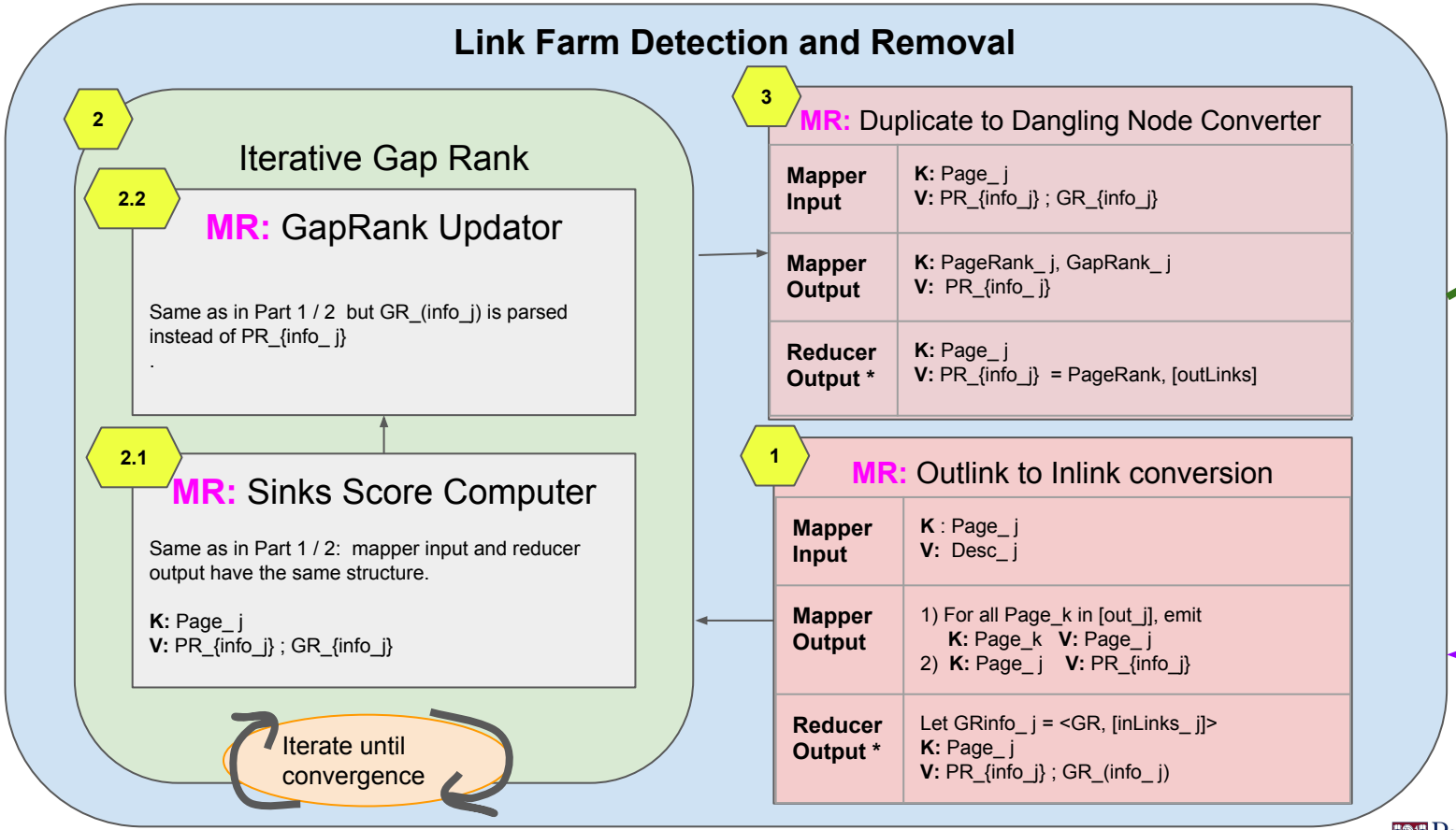
Emit
(<url>, <title>\t<description>)

IV. Page Rank

PageRank Architecture Part 1/2



PageRank Architecture Part 2/2



*Note this is the same output format as input to PageRank phase)

Let **info_ j** = <Rank_ j, [inLinks_ j] > when used as a subscript for GR.

V. Search Engine

Search Engine Architecture