

# *junosearch* Architecture Diagrams

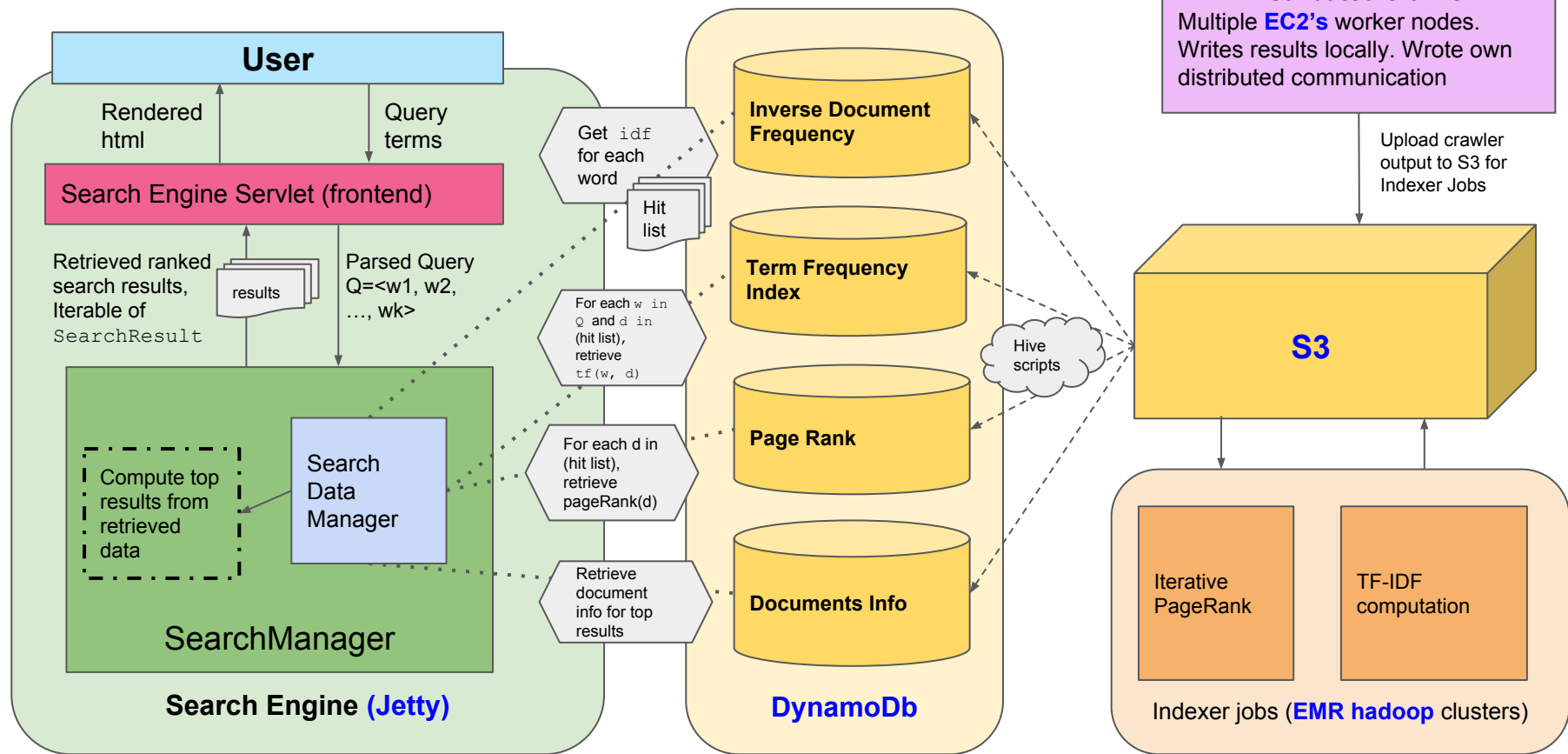
University of Pennsylvania, CIS, School of Engineering and Applied Sciences  
CIS 455 / 555 Final Project

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)

junosearch

---

buy stash

---

SEARCH

I'M FEELING LUCKY

# Front End Samples Demo (results page)

junosearch

## Ebay Search Results



Sentry Safe Fireproof Fire Chest Security Lock Money Document Stash Gun Box NEW

Price: 25.99

SENTRY SAFE FIREPROOF  
FIRE CHEST SECURITY  
LOCK MONEY DOCUMENT  
STASH GUN BOX NEW



2x Airtight Smell Proof Container - Aluminum Herb Stash Jar

Price: 14.59

2X AIRTIGHT SMELL  
PROOF CONTAINER -  
ALUMINUM HERB STASH  
JAR



Safe Fireproof Fire Chest Security Lock Money Document Stash Gun Box 0.18 Cu New

Price: 21.92

SAFE FIREPROOF FIRE  
CHEST SECURITY LOCK  
MONEY DOCUMENT STASH  
GUN BOX 0.18 CU NEW



Arizona Green Tea Diversion Safe Can Stash Arizona

Price: 17.28

ARIZONA GREEN TEA  
DIVERSION SAFE CAN  
STASH ARIZONA

## Git - git-config Documentation

<https://git-scm.com/docs/git-config#git-config-tarumask>

Rank Score:3.334539603037935E-4

## Default Title

[https://www.w3schools.com/browsers/browsers\\_opera.asp](https://www.w3schools.com/browsers/browsers_opera.asp)

The was no description for this page

Rank Score:2.1622266489069498E-4

## Default Title

<https://git-scm.com/docs/git-config/2.3.0>

The was no description for this page

Rank Score:1.6672698015189675E-4

## Default Title

<http://www.thesaurus.com/browse/saved>

The was no description for this page

Rank Score:1.563065438024032E-4

## Wikipedia Search Results

### Stash

<https://en.wikipedia.org/wiki/Stash>

A stash is a large personal collection that is often kept secret. Stash also may refer to: Stash Hotel Rewards Stash Tea Company Stash (band), a Belgian

### Stash Tea Company

[https://en.wikipedia.org/wiki/Stash\\_Tea\\_Company](https://en.wikipedia.org/wiki/Stash_Tea_Company)

Stash Tea Company is a privately held specialty tea & herbal tea company headquartered in Tigard, Oregon, a suburb of Portland. Stash Tea was founded

### Stash Box

[https://en.wikipedia.org/wiki/Stash\\_Box](https://en.wikipedia.org/wiki/Stash_Box)

Stash Box is the second EP-CD from the Kottonmouth Kings released only in Japan to support the upcoming Japanese tour. The CD was released on March 10

### Hidden Stash II: The Kream of the Krop

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

**Term\_Frequency\_Index** (stores tf (w, d) for each w,d pair)

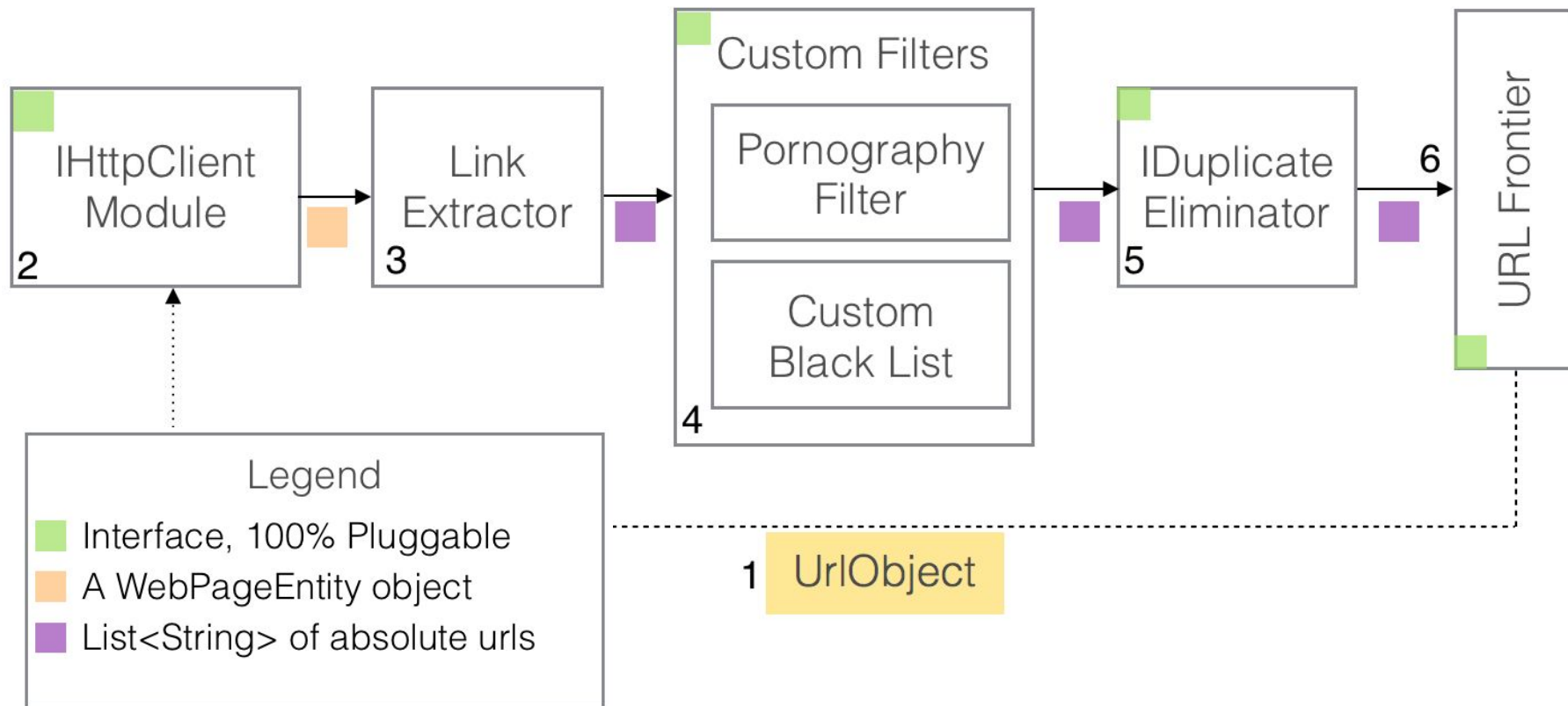
```
Term {
    "term": "String"; // primary partition key
    "tf": "Number"; // primary sort key

    // 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>";
}
```

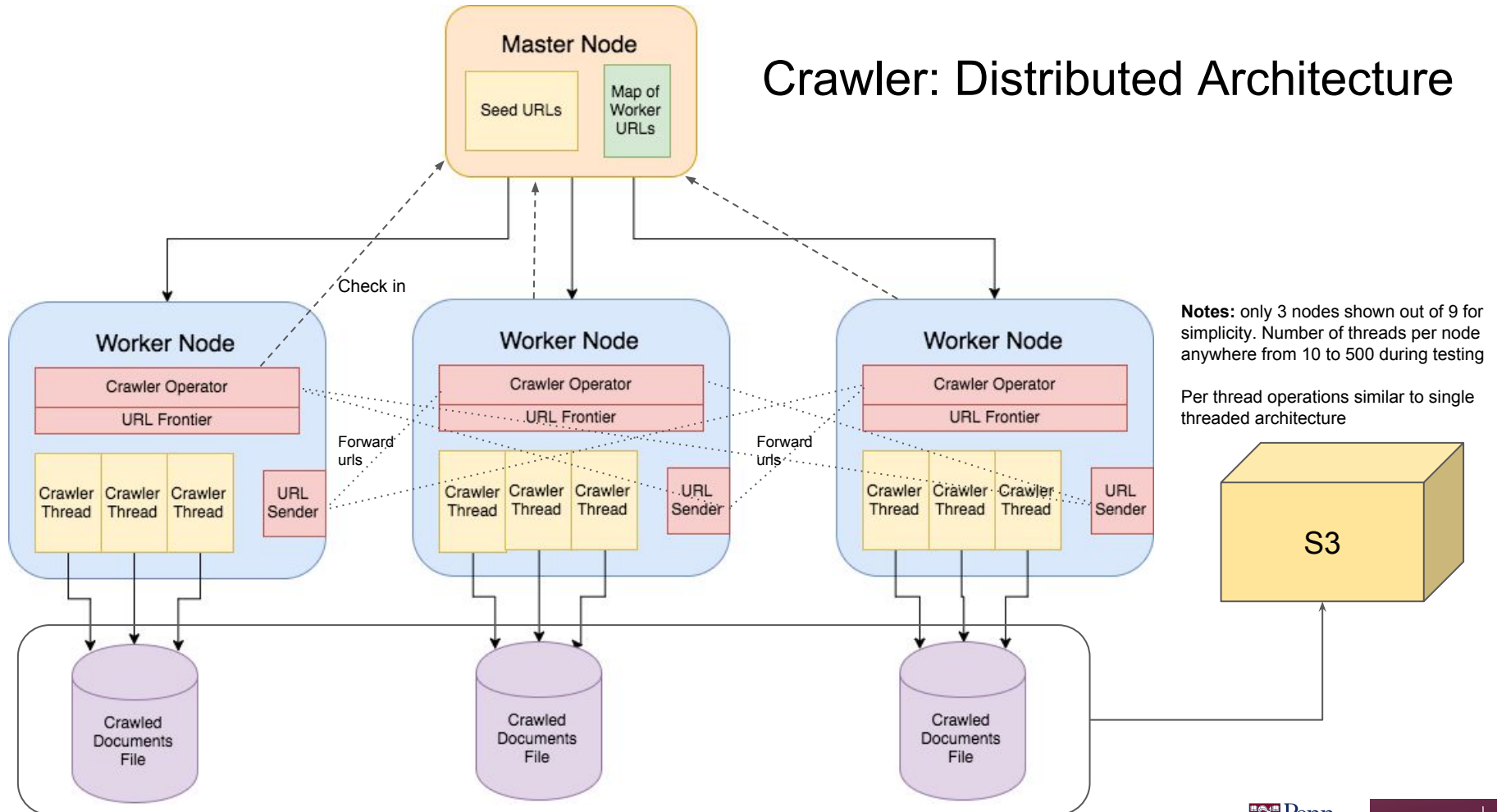
## **II. Crawler Diagrams**

# Crawler: Single Threaded, Single Node Design





# Crawler: Distributed Architecture



# **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>

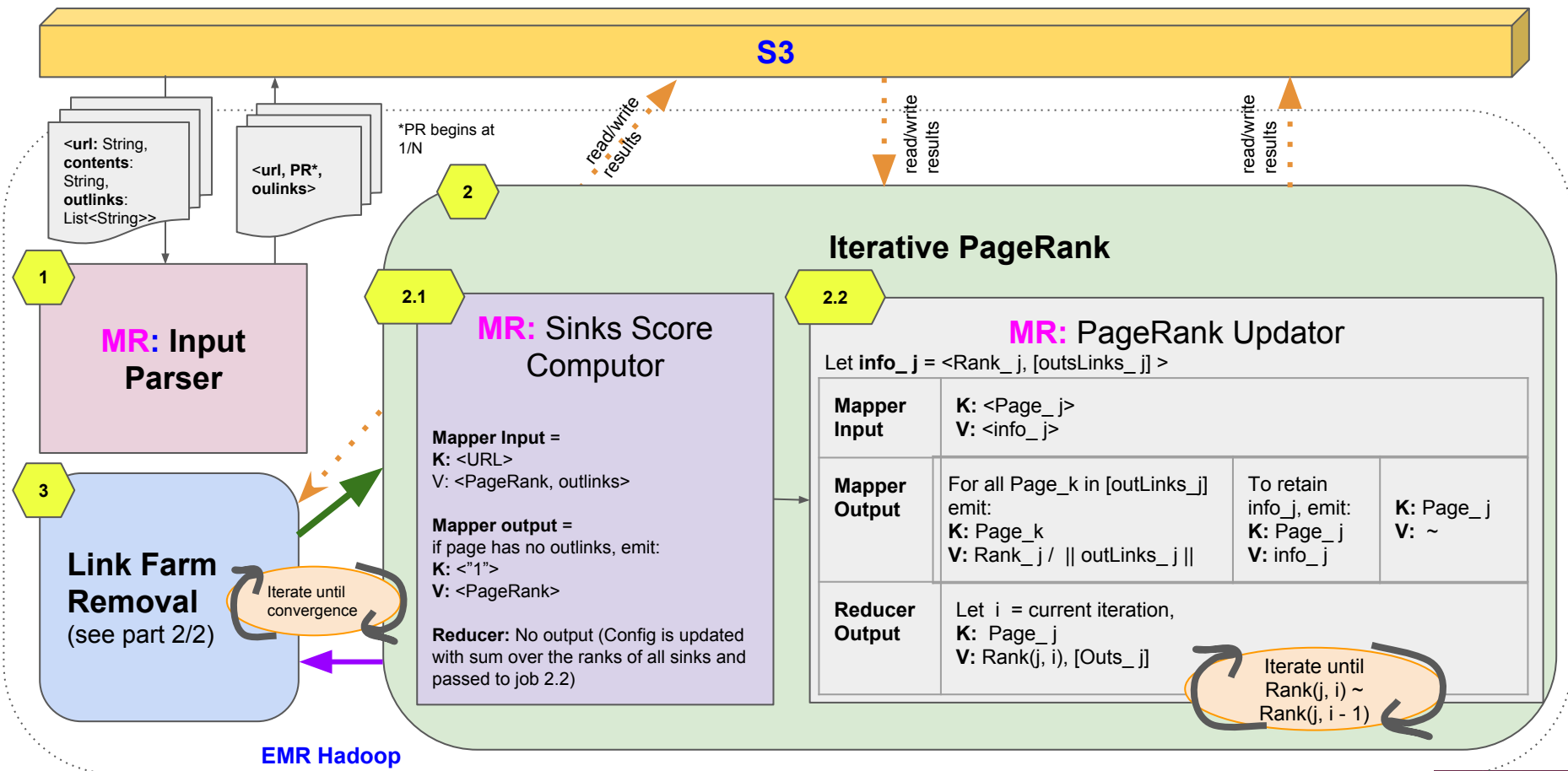
**Value** <title>\t<description>

### Emit

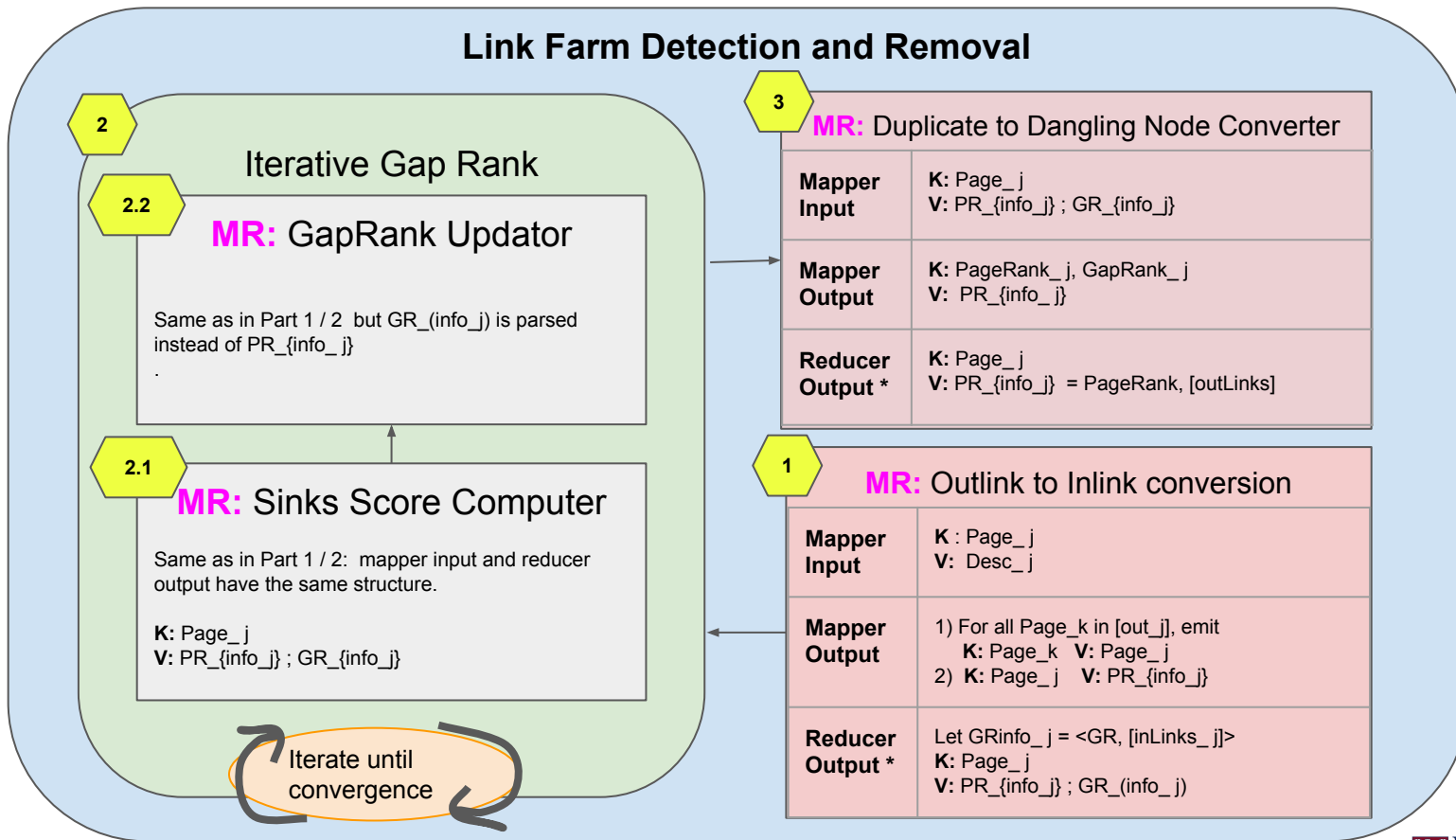
(<url>, <title>\t<description>)

## IV. Page Rank

# PageRank Architecture Part 1/2



## Link Farm Detection and Removal



\*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

