



Bixo - a webcrawler toolkit

Ken Krugler, Stefan Groschupf

Agenda

Overview

Background

Motivation

Goals

Status

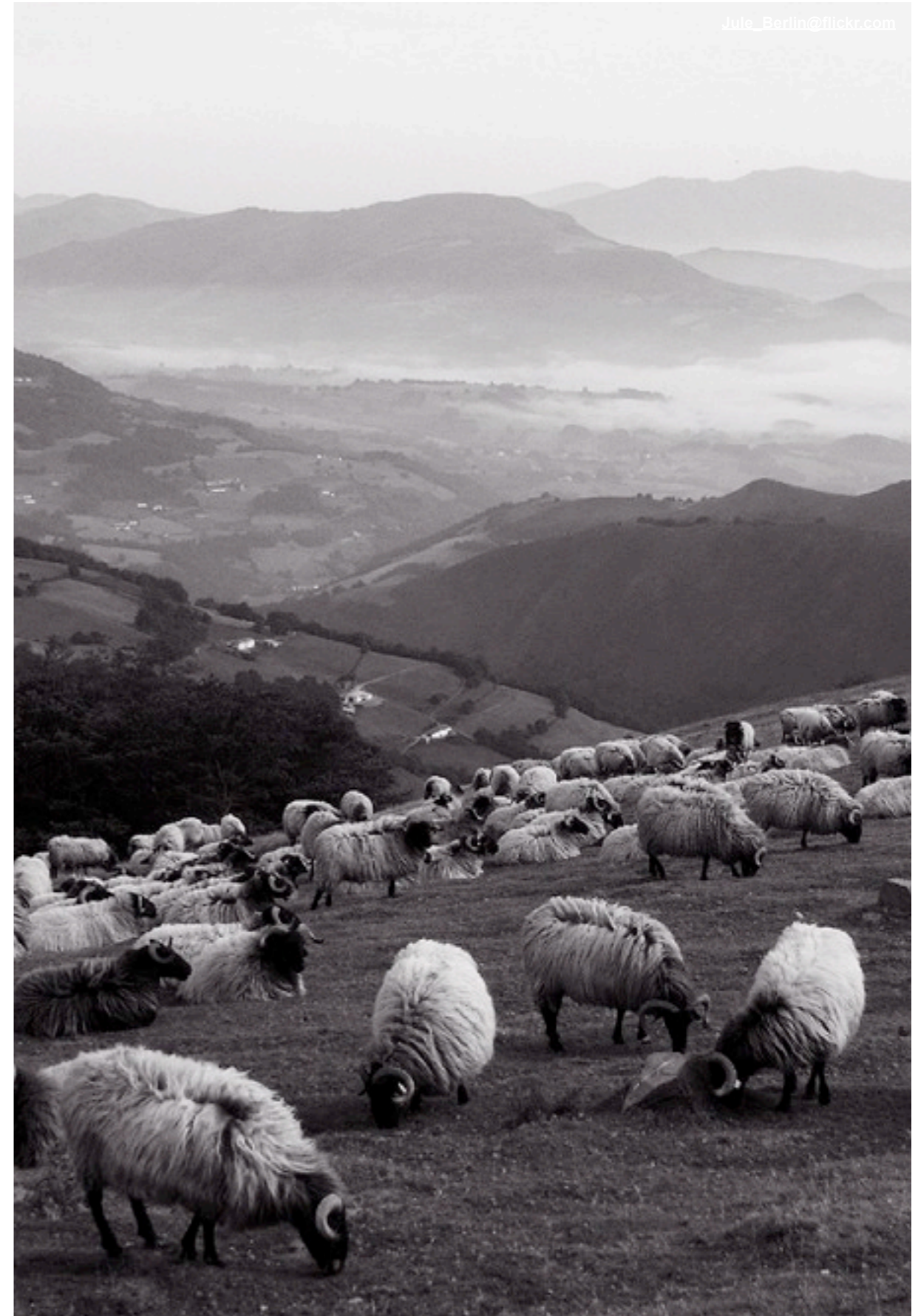
Differences

Architecture

Data life cycle

Robust Testing

Resources



Overview

- ▶ Primary users will be companies extracting data from the web (not search)
- ▶ Interested in subset of the web
- ▶ Typically part of larger data processing system

Motivation - tech

- ▶ No good solution available
- ▶ We need a **toolkit**
- ▶ Missing from Nutch et al.
 - ▶ Easy to integrate
 - ▶ Easy to extend
 - ▶ Easy to understand
 - ▶ API vs CLI
 - ▶ Pluggable I/O
- ▶ Avoid common problems
 - ▶ Spider traps & link farms
 - ▶ Slow servers
 - ▶ Hanging crawls

Motivation - EMI

- ▶ Screen scrape, data extraction
- ▶ Artist websites, e.g. concert dates
- ▶ Many pages from large sites
- ▶ Just crawl, no index
- ▶ One of many inputs into Business Intelligence
- ▶ Integration in larger BI system (Cascading-based)

Motivation - Share This

- ▶ Focused index for key partners
- ▶ Data analysis and mining of 100m pages
- ▶ Integration into existing log analysis and data mining systems (Cascading-based)
- ▶ Low IT/Ops support requirements

Goals

- ▶ Fulfill key motivating requirements
- ▶ OSS project with business-friendly license
- ▶ Focus on vertical crawling, leverage other projects
- ▶ Efficient execution in EC2/cloud environment
- ▶ Grow OSS community

Current Status

- ▶ We already do crawls in EC2
- ▶ 2 sponsored developers, since March 2009
- ▶ MIT license
- ▶ Todo:
 - ▶ Improve robots.txt handling
 - ▶ Bugfixes and many improvements
 - ▶ Website & documentation
 - ▶ A CLI for easy testing.

Differences (from Nutch)

- ▶ Toolkit versus system - building blocks, not plugins
- ▶ Workflow focus, versus system where you set conf and run a command
- ▶ More emphasis on instrumentation - monitoring, error handling,
- ▶ No search serving
- ▶ Vertical crawl, not intranet or whole web
- ▶ HTTP(S) only, not ftp, etc.

Differences (from Hadoop)

- ▶ Not much, which is a good thing
- ▶ Generates lots of data - want to store in S3, want to minimize writes
- ▶ Heavy user of DNS server - extra set up for caching server
- ▶ Fetch phase is unusual Cascading topology

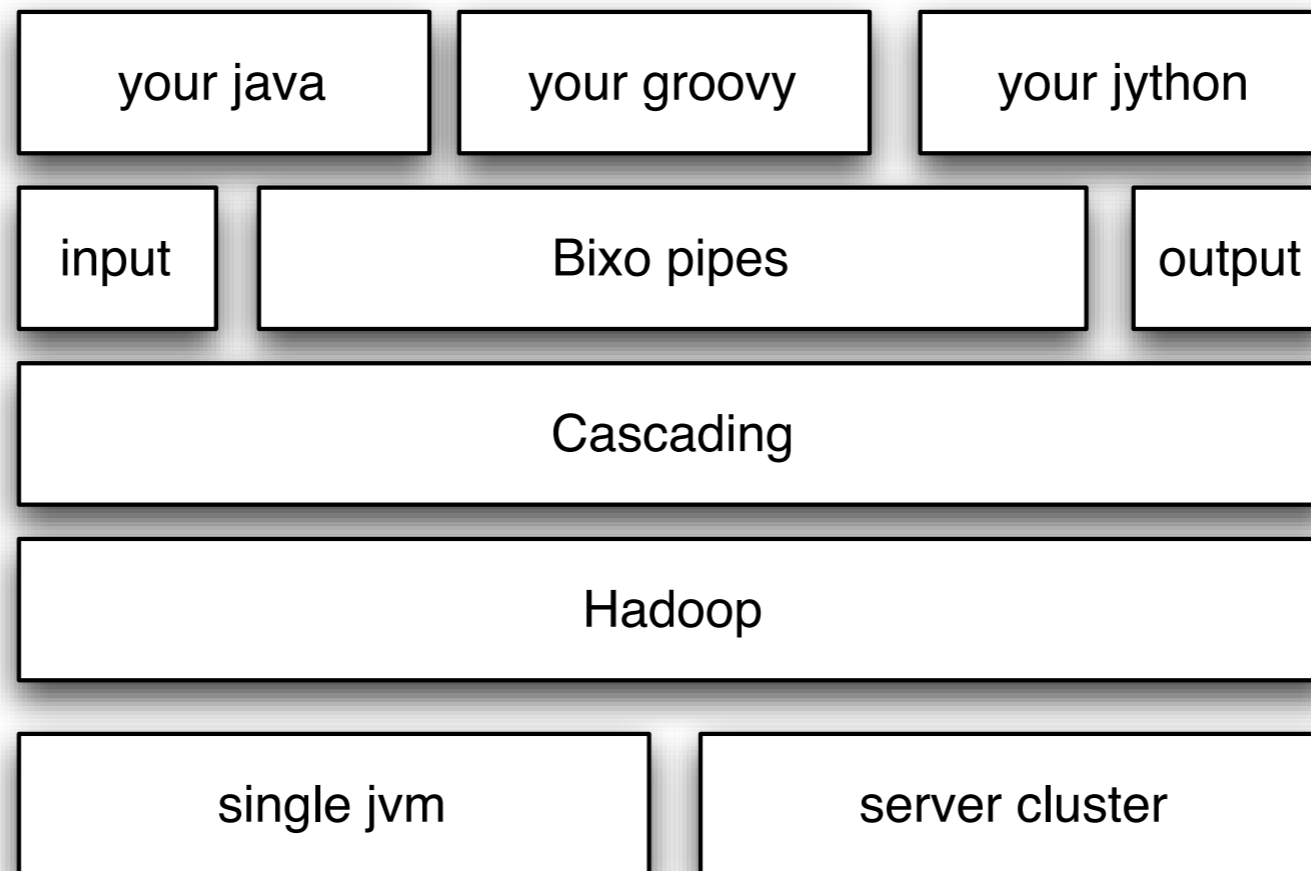
Hadoop Intro

- ▶ Open Source map reduce system
- ▶ Execution layer - map reduce
 - ▶ Mapper, Reducer Tasks
- ▶ Storage layer - (distributed) file system
 - ▶ Local FS, HDFS, S3, etc
- ▶ Scales from single node to thousands

Cascading Intro

- ▶ Data processing can be hard with Hadoop
- ▶ Cascading extends Hadoop
- ▶ Provides simple data processing API
- ▶ Reusable (unix) pipe based concept
- ▶ Sources and Sinks separated
 - ▶ HDFS, Hbase, JDBC, Aster etc.
- ▶ Assemble Pipes, Source and Sink in a Flow
- ▶ GPL or OEM, though might change

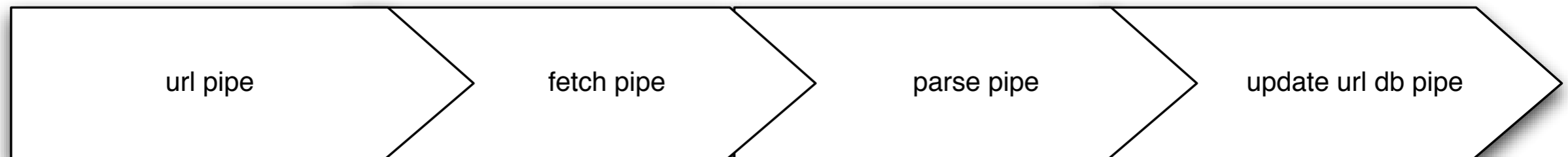
Architecture



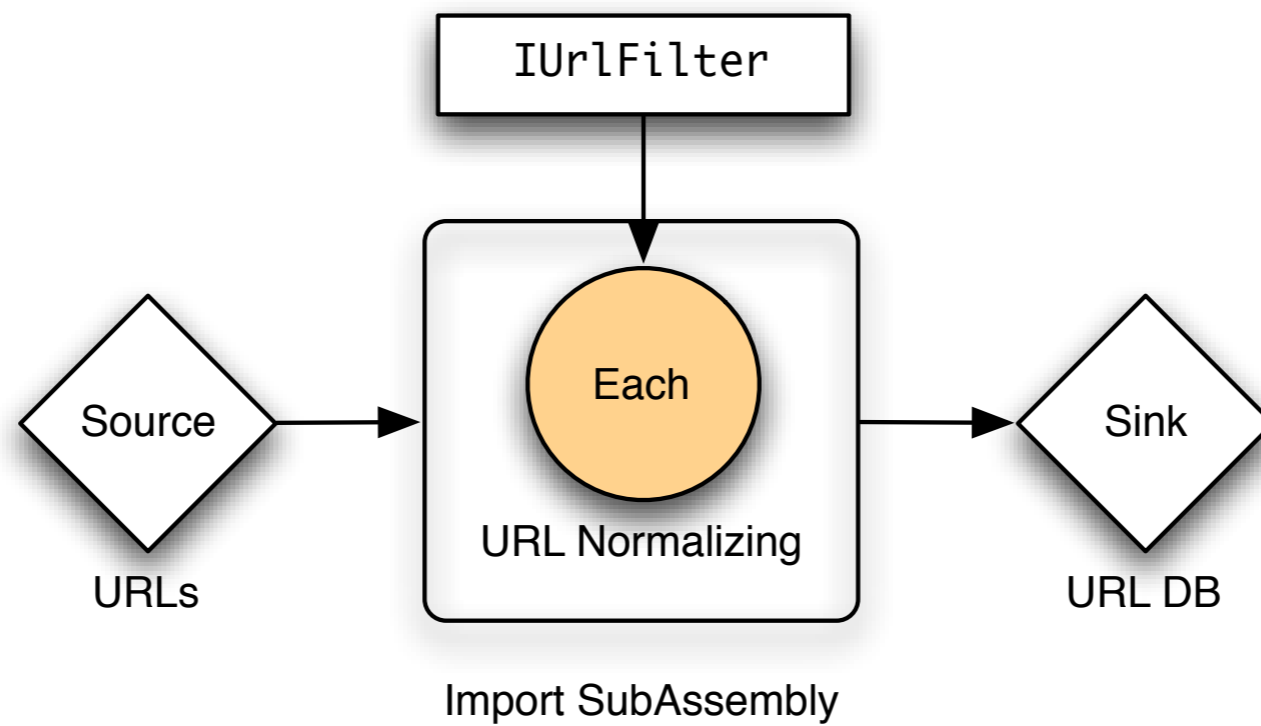
Data life cycle

- ▶ Inject URLs in URL DB
- ▶ Select URLs from URL DB - based on recrawl policy, or partner/domain, or type, etc
- ▶ Normalize URLs
- ▶ Score URLs
- ▶ Group URLs
- ▶ Fetch
- ▶ Save content
- ▶ and/or update URL DB
- ▶ and/or analyze/parse content
- ▶ Notice nothing about indexing, pushing out index, serving up index.
- ▶ Meta data fully supported

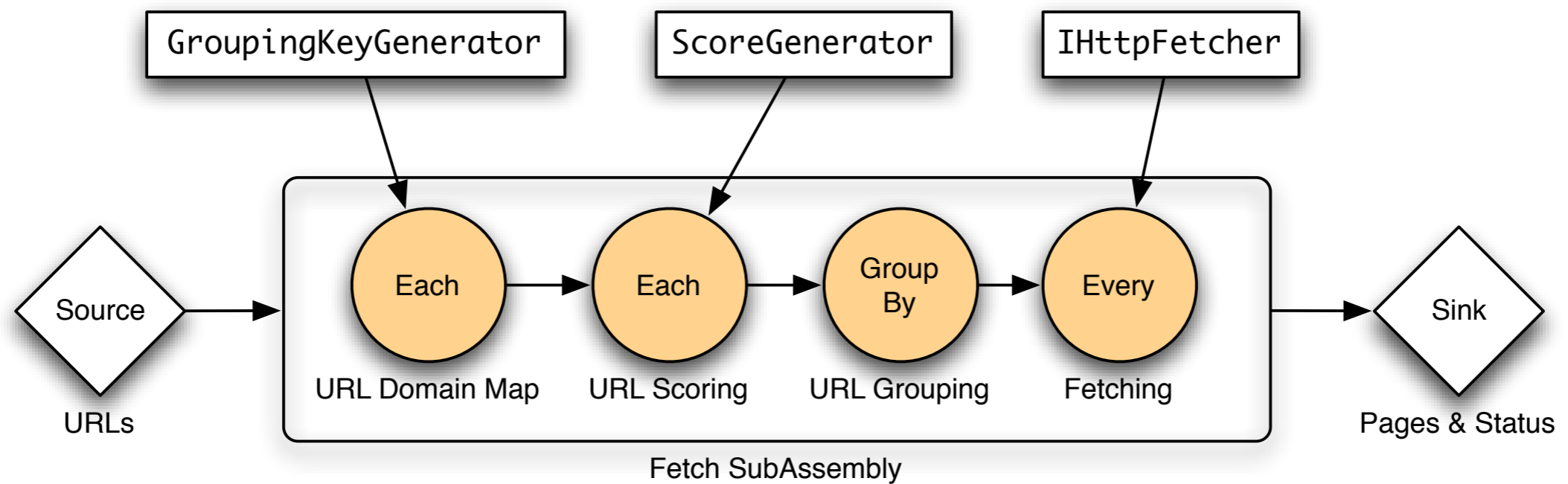
Architecture - Pipes



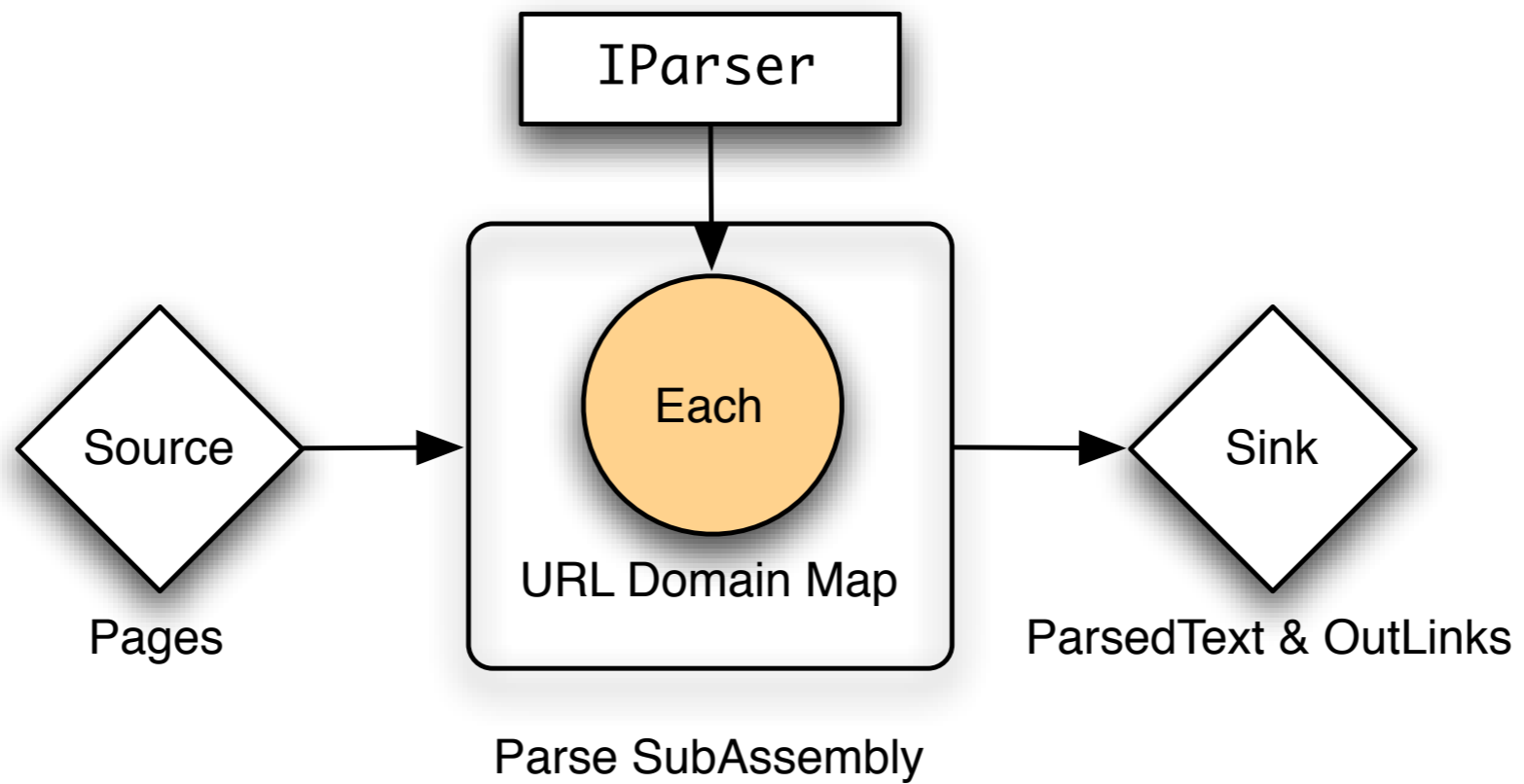
Import Url Pipe



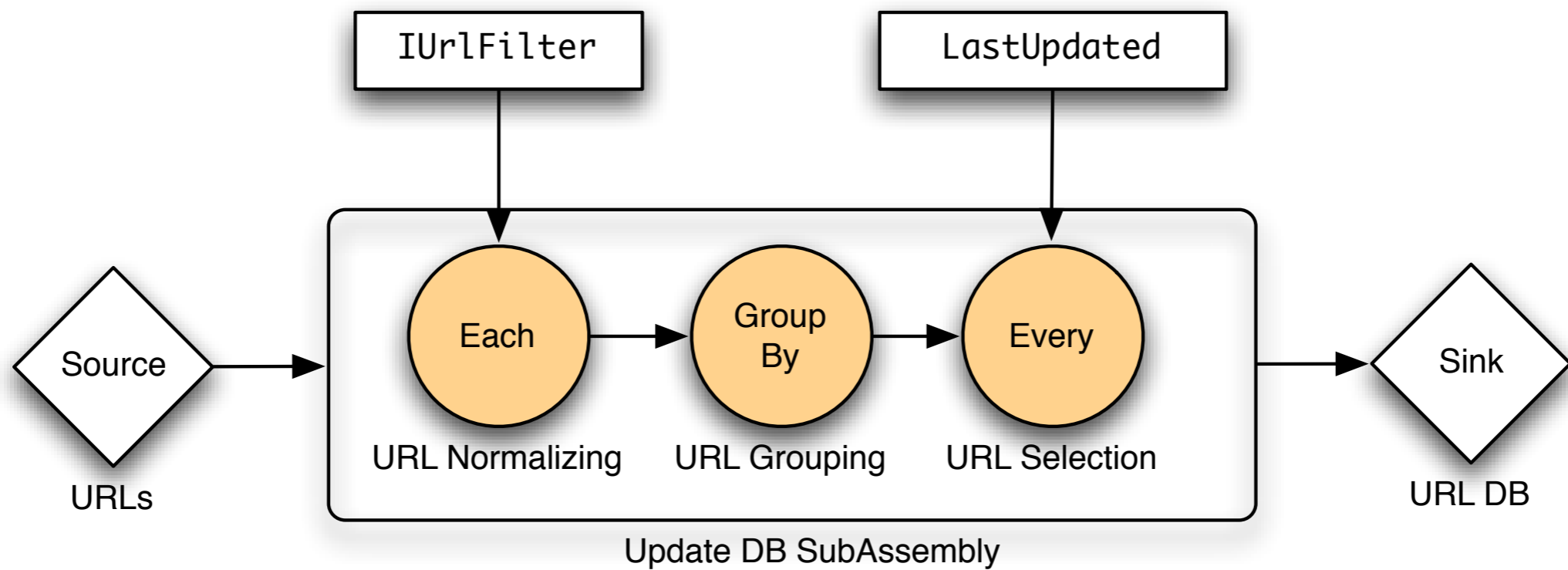
Fetch Pipe



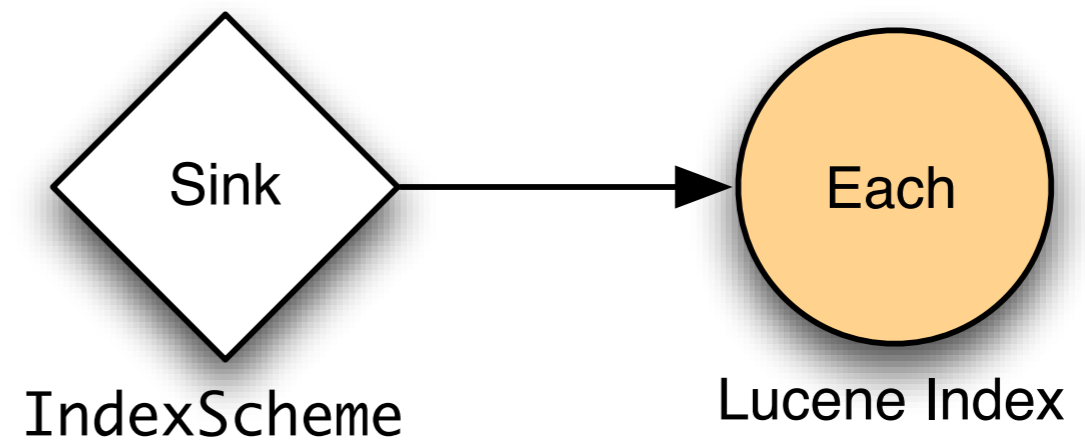
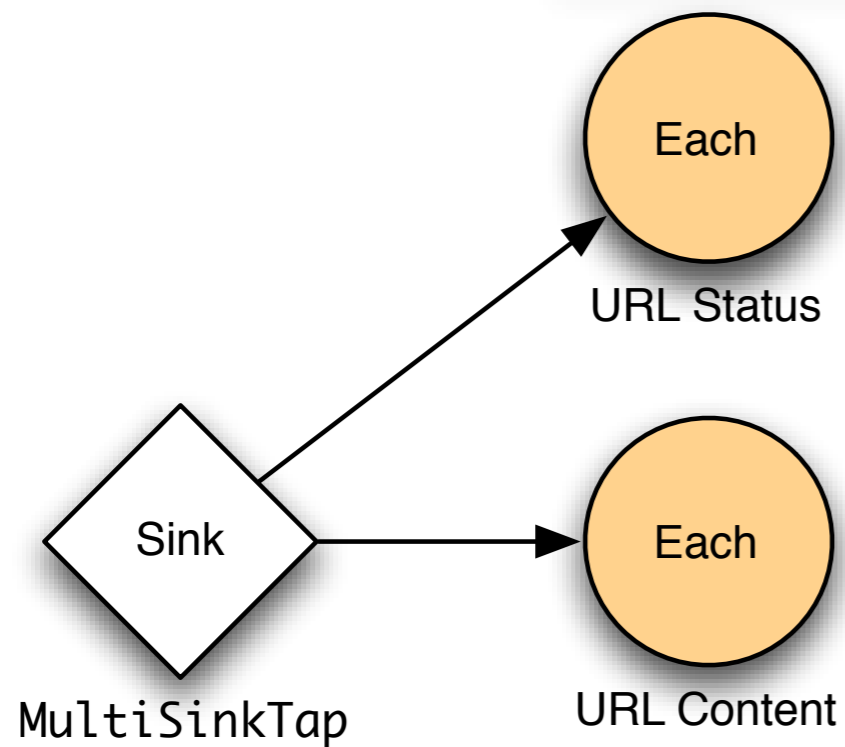
Parse Pipe



Update Pipe



Output



Robust testing

- ▶ Unit tests
- ▶ Jetty with special request handlers
 - ▶ wrong content type
 - ▶ slow responses
 - ▶ wrong header
- ▶ WebGraph test platform
 - ▶ test/simulate URL discovery
 - ▶ Looping/URL DB updates
 - ▶ page rank calcs, etc.
- ▶ Wikipedia
 - ▶ large amount of data that can be "crawled" via local setup

<http://webgraph.dsi.unimi.it/>

Resources

Web:	http://bixo.l0ltec.com/
List:	http://groups.yahoo.com/group/bixo-dev
Sources:	https://github.com/emi/bixo/tree
Bug tracking:	http://oss.l0ltec.com/jira/browse/bixo



Scale Unlimited, Inc.

Ken Krugler, Stefan Groschupf

info@scaleunlimited.com