# Large Scale Text Analysis Using the Map/Reduce Hierarchy

**David Buttler**

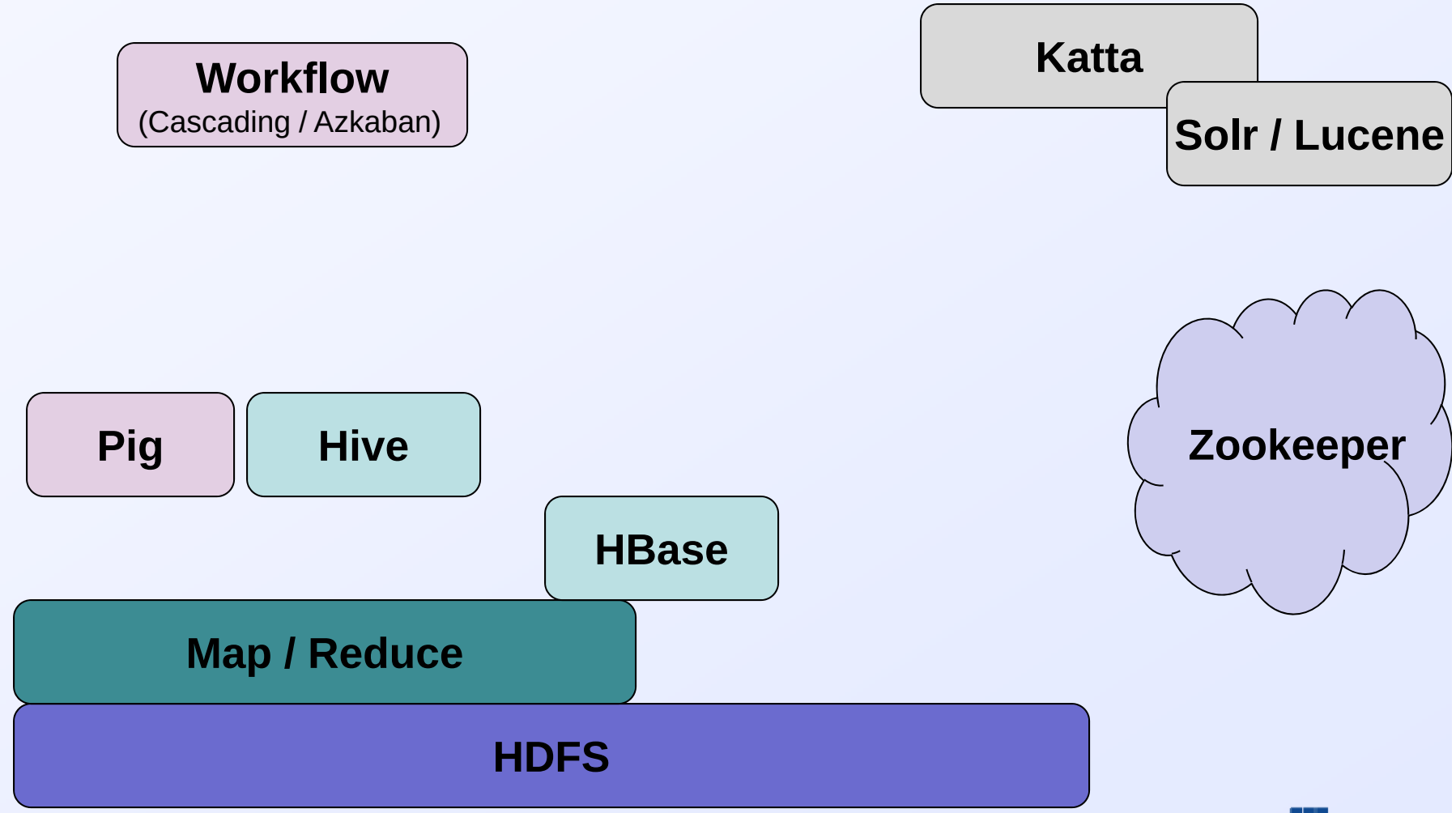
This work is performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

# Large scale computing with commodity hardware

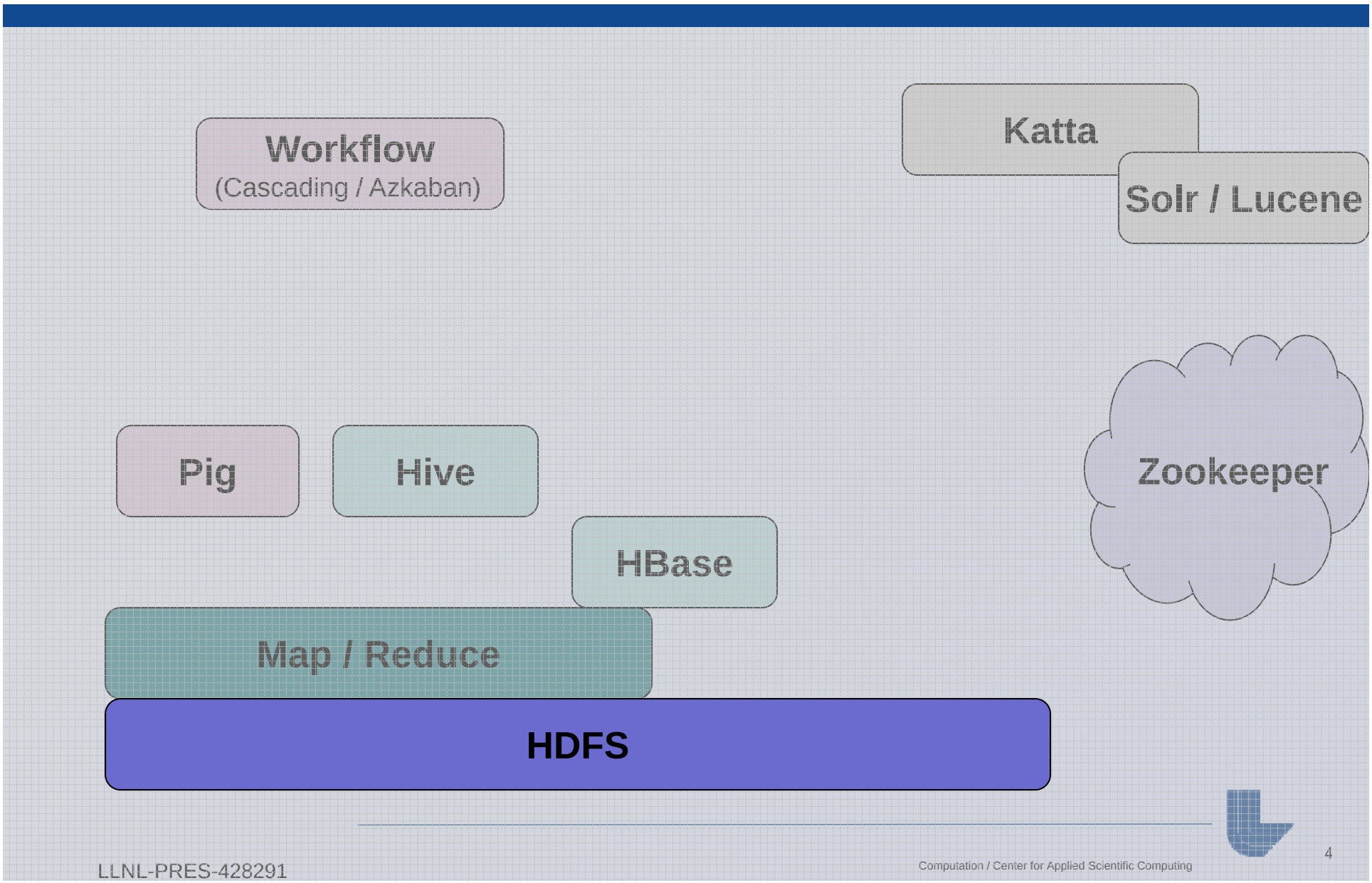
- Origins
- Google GFS, Map/Reduce, BigTable
- Microsoft Azure
- **Hadoop**: Yahoo! / Open Source Software

- Why do we care: k-mer lexing
  - 10 hours on a single fat node
  - 1 hour on an old cluster

# The M/R stack of open source software

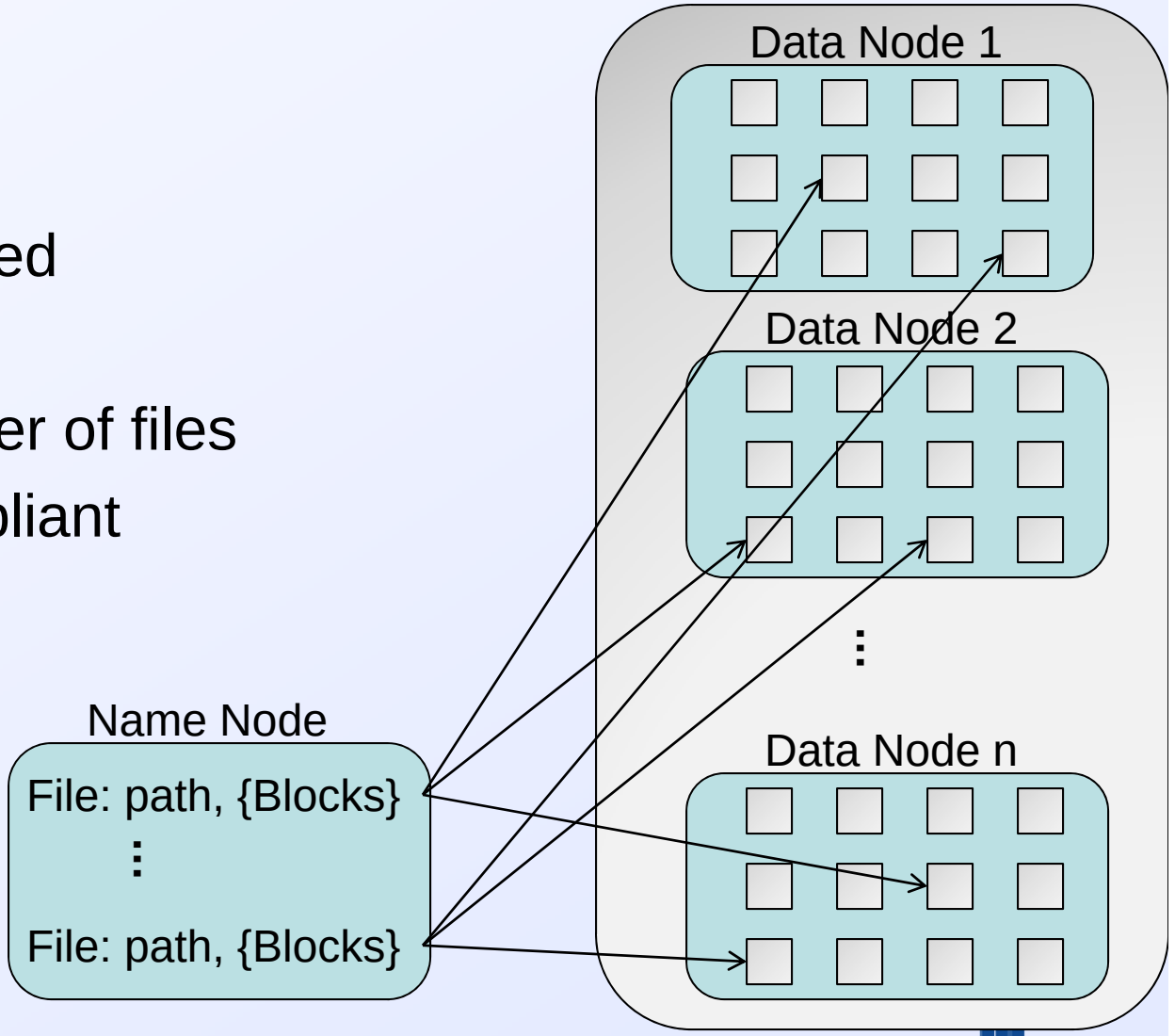


# The M/R stack of open source software – HDFS

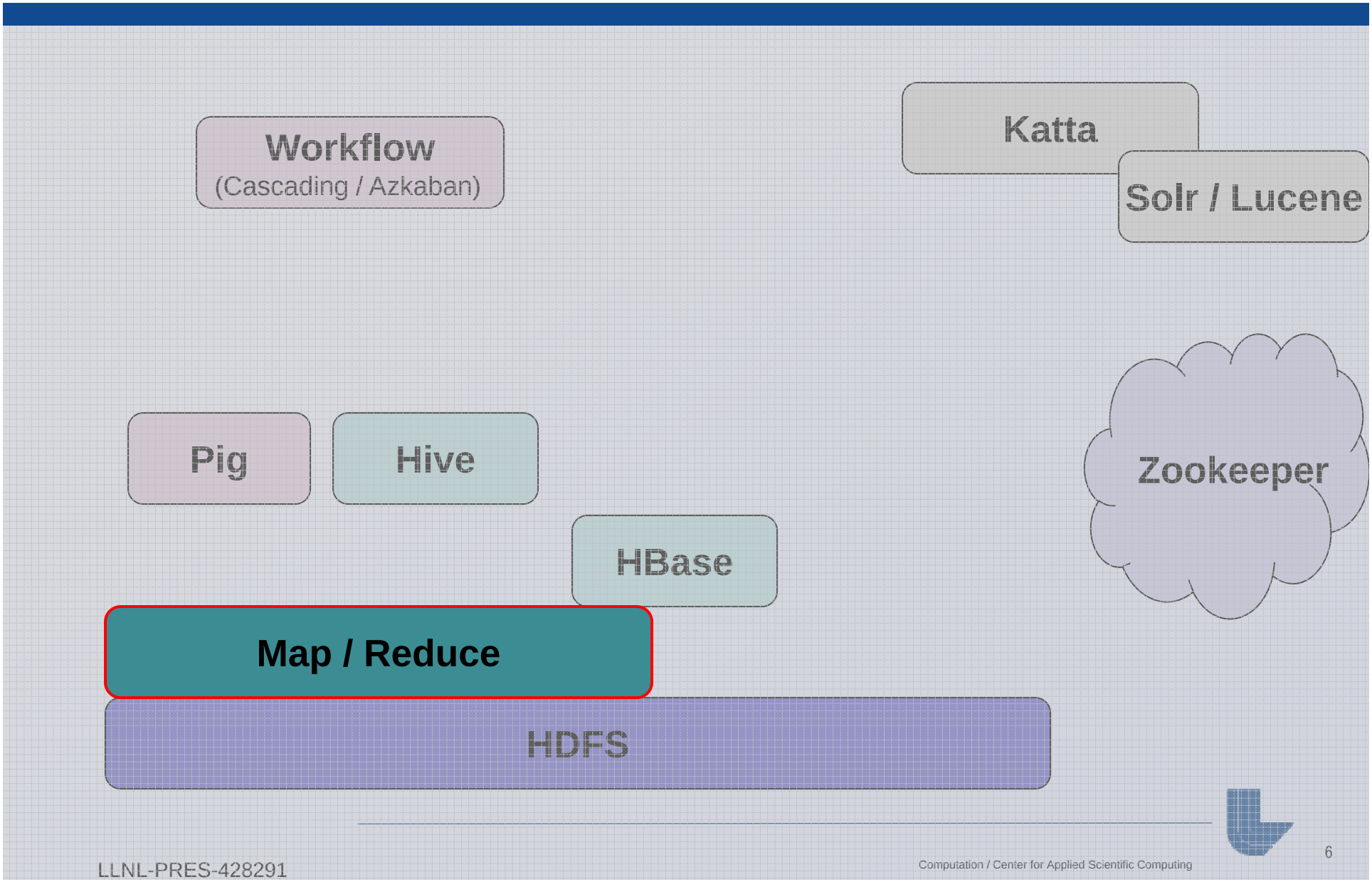


# HDFS

- Replicated
- Distributed
- Centrally managed
  - SPOF
  - Limited number of files
- Not POSIX compliant
- Rack-aware

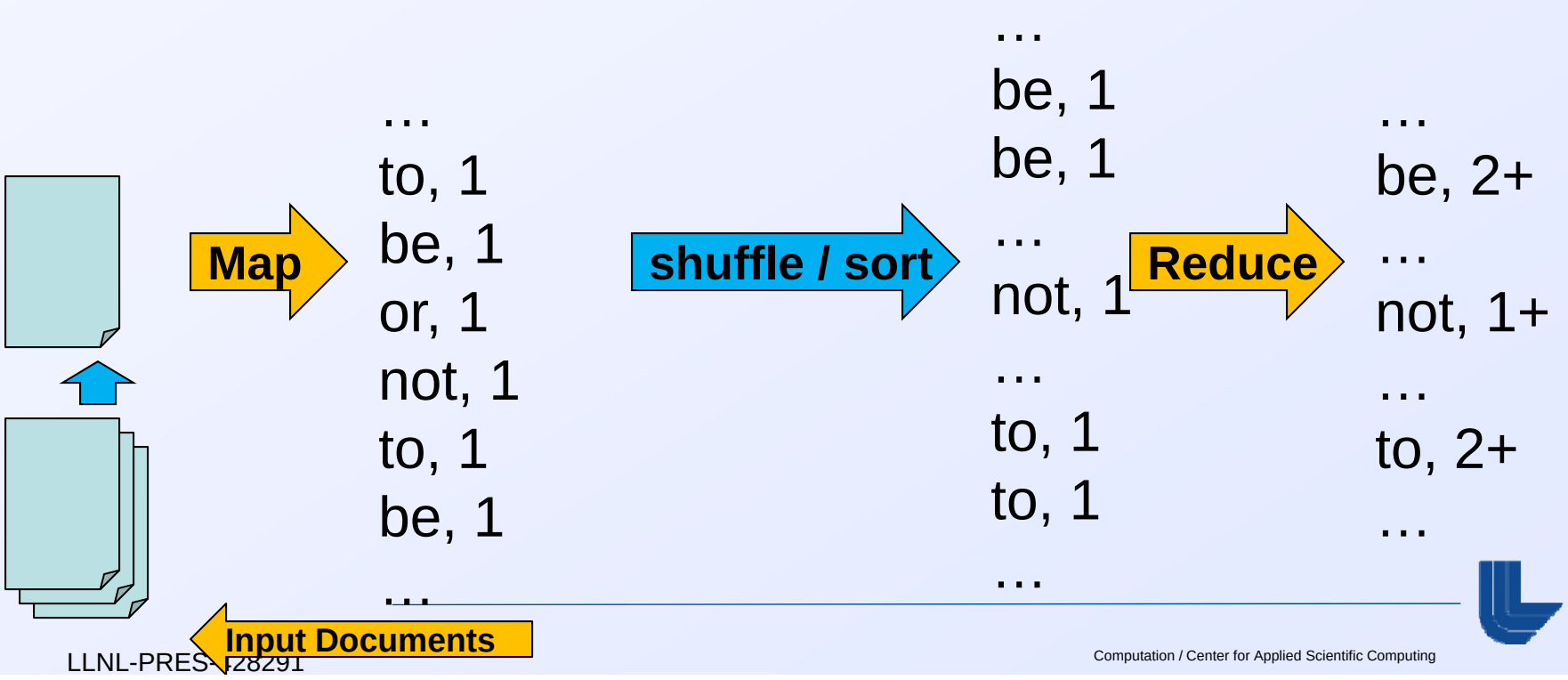


# The M/R stack of open source software – Map / Reduce



# Map/Reduce is functional programming distributed over a cluster

- Distributed computation
- Two phase computation
- Built-in shuffle/sort between phases
- Canonical example: word frequency count for the web

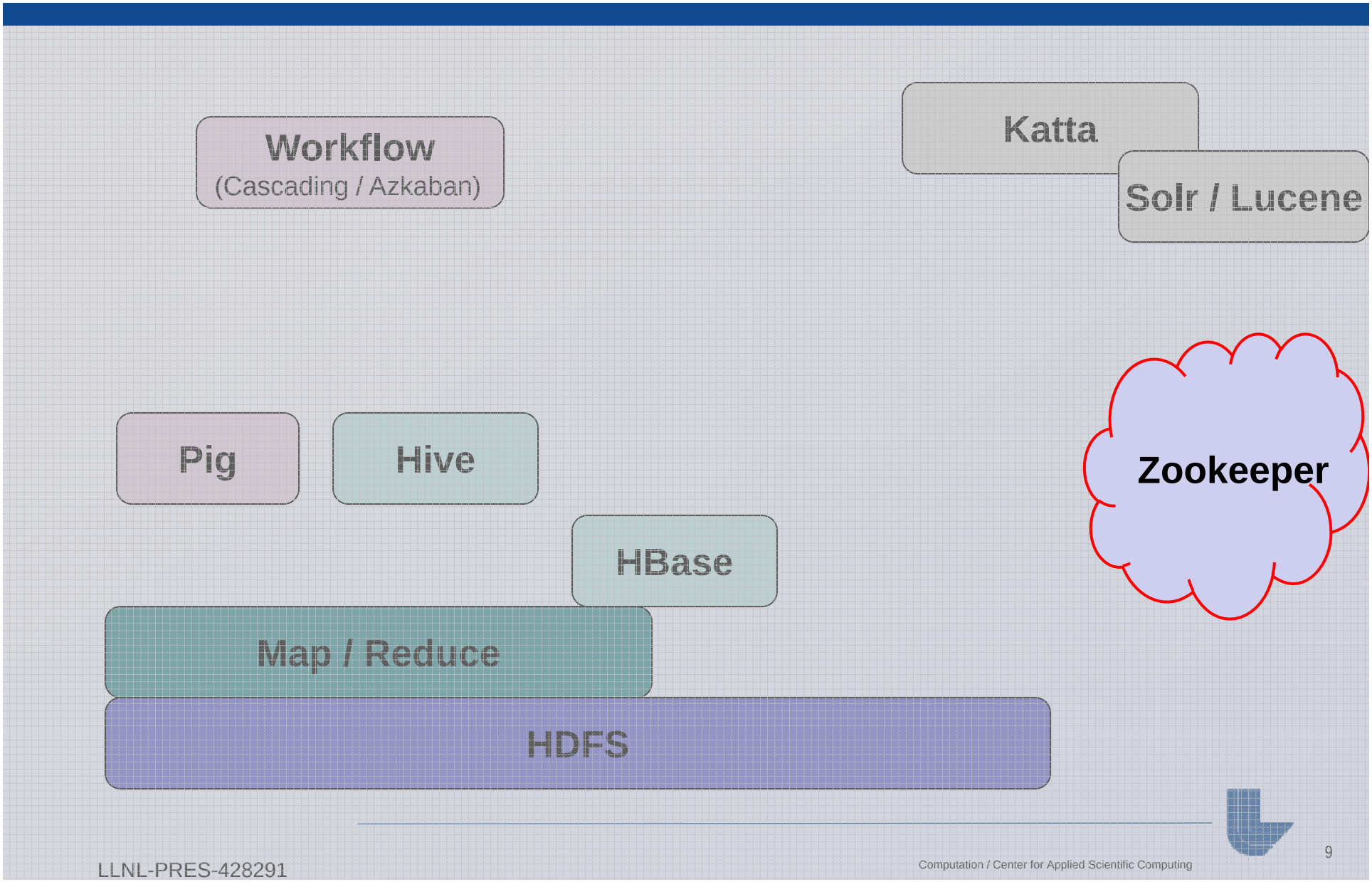


# More interesting M/R examples

- Map input: document
- Map output: raw text

- Map input: text
- Map output: Named entity annotations

# The M/R stack of open source software – Zookeeper



# Zookeeper

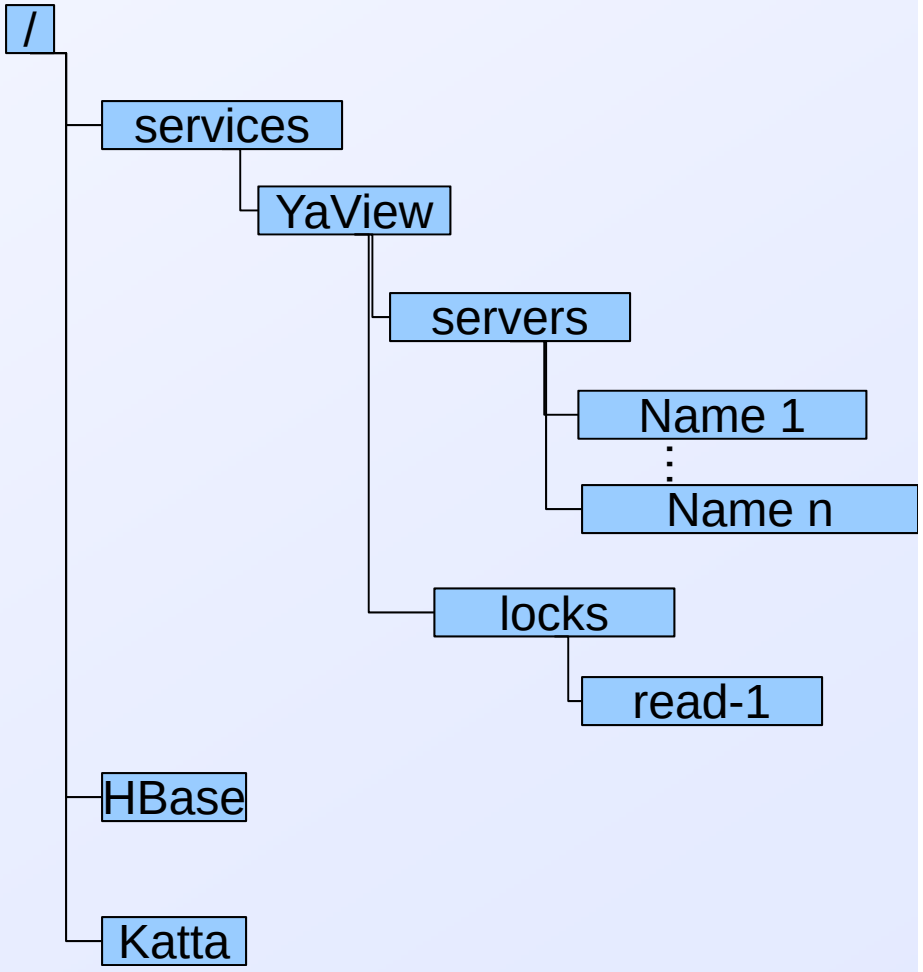
- A highly available, scalable, distributed, configuration, consensus, group membership, leader election, naming, and coordination service
- Uses:
  - HBase: row locking; region key ranges; region server addresses
  - Katta: shard location information
  - Message queues
- Not: a large scale data store

# Zookeeper Guarantees

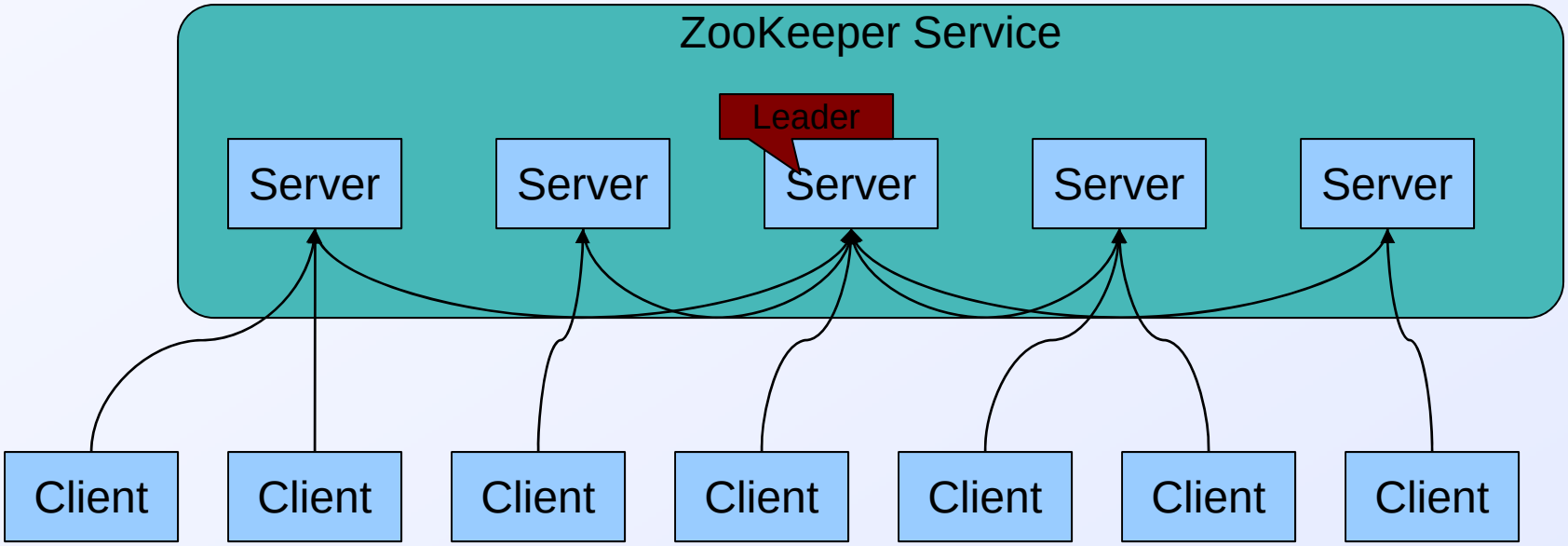
1. Clients will never detect old data.
2. Clients will get notified of a change to data they are watching within a bounded period of time.
3. All requests from a client will be processed in order.
4. All results received by a client will be consistent with results received by all other clients.

# Zookeeper Data Model

- Hierarchal namespace
- Each znode has data and children
- data is read and written in its entirety
- Nodes store < 1MB data
- Writes go to all nodes

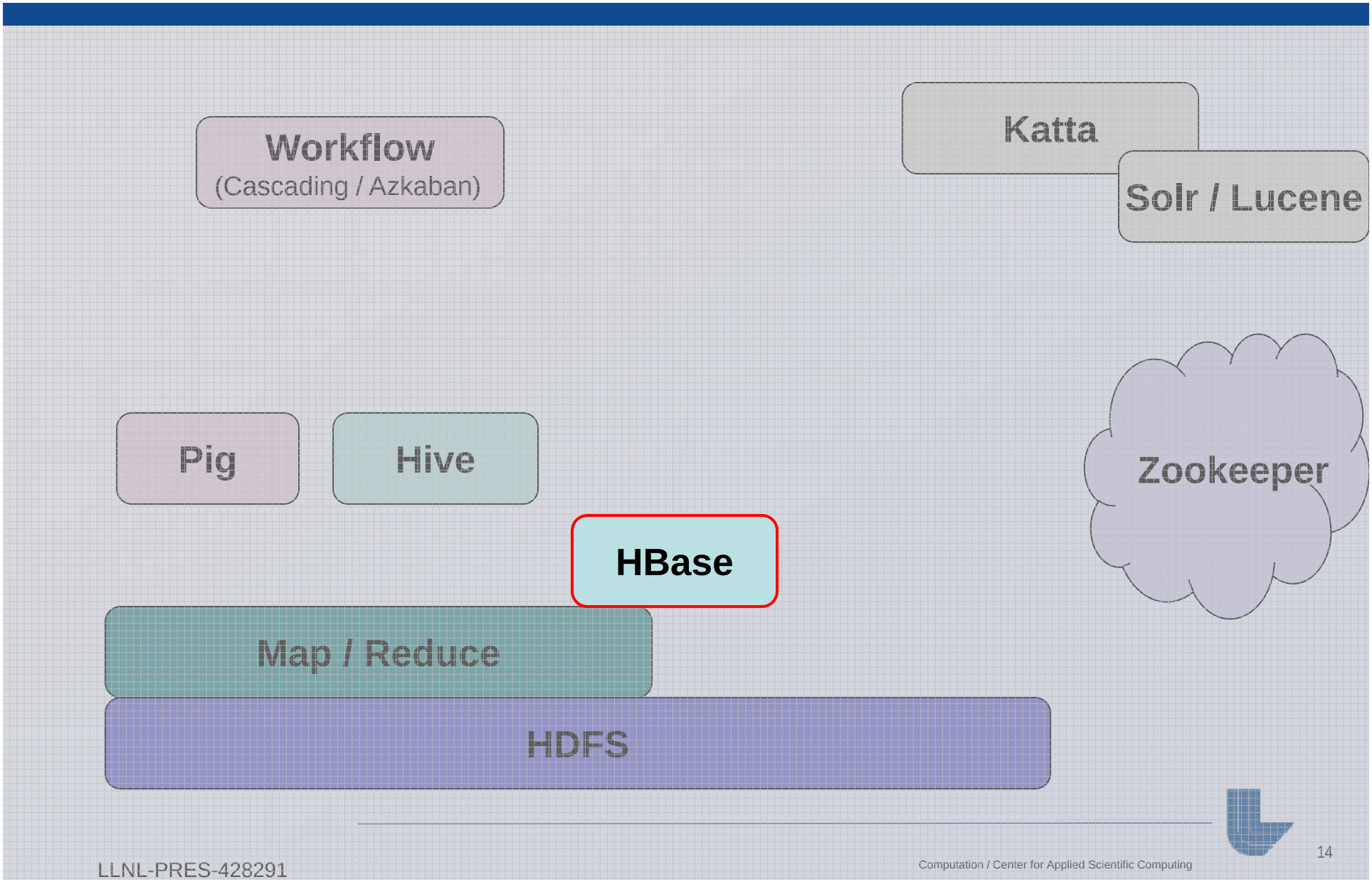



# ZooKeeper Service




- All servers store a copy of the data (in memory)
- A leader is elected at startup
- Followers service clients, all updates go through leader
- Update responses are sent when a majority of servers have persisted the change

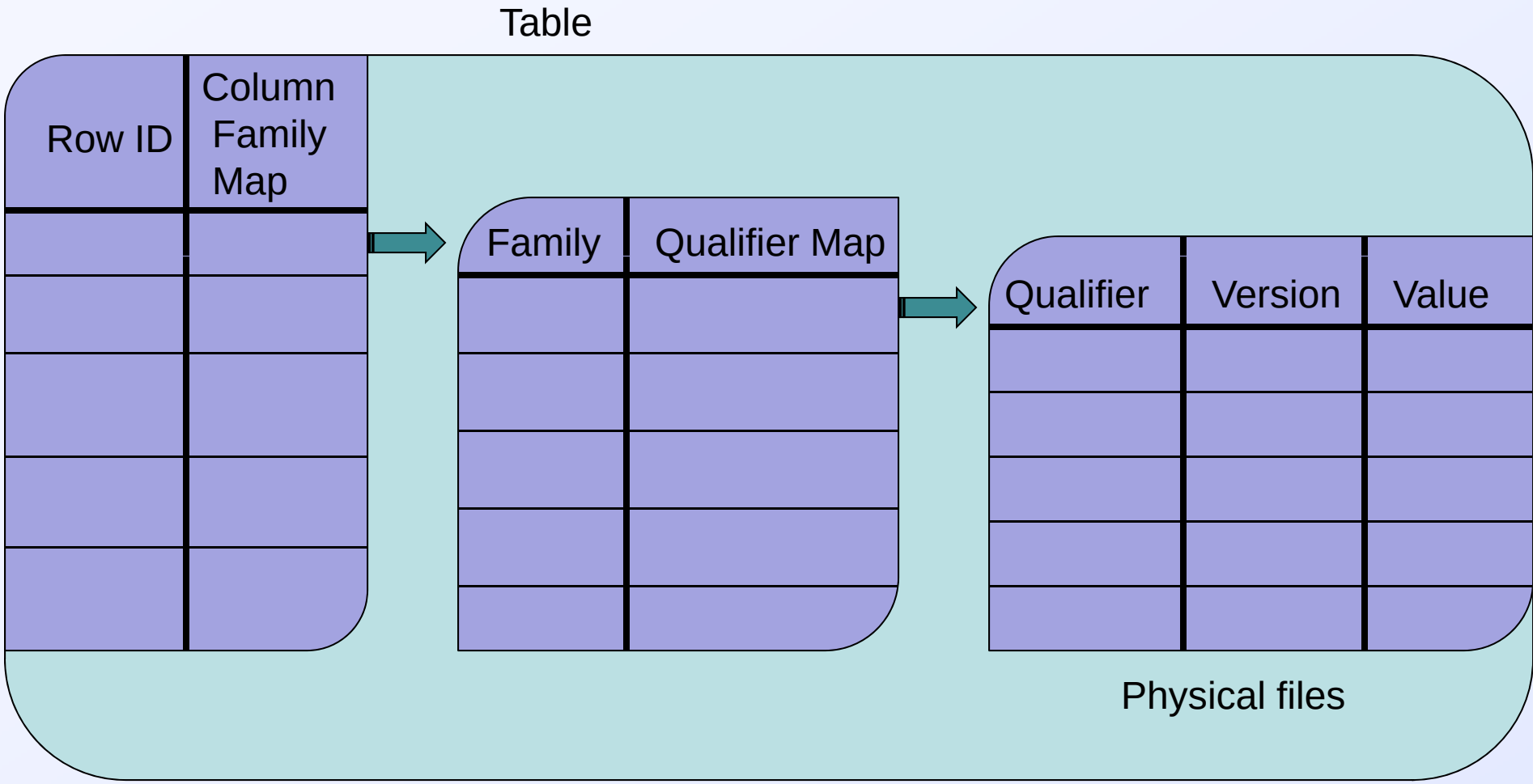
# The M/R stack of open source software – HBase



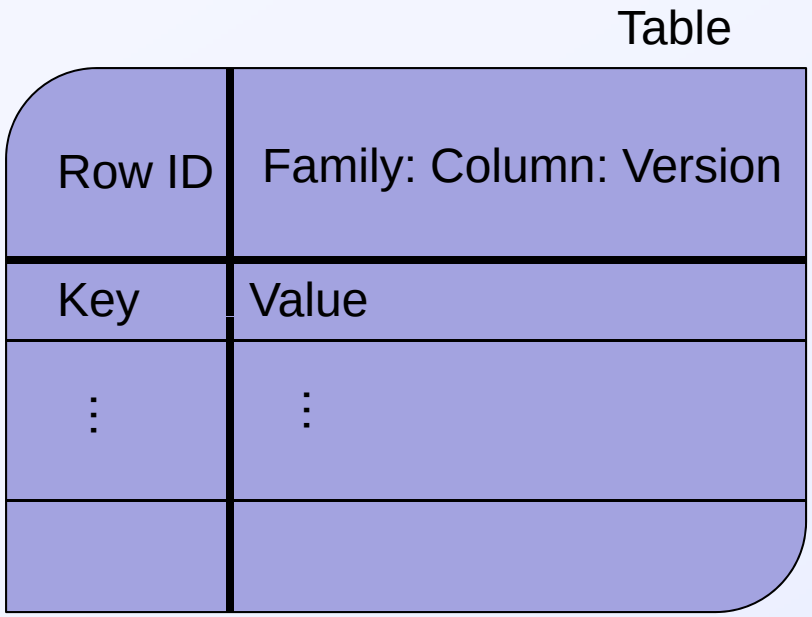
# HBase

- Distributed column oriented data store
  - Only supports one data type
  - Tables are broken into regions
  - Regions are automatically split and redistributed
  - All data is local
- Scales to > 1M row / second insert rate (20 node cluster)
- Tightly integrated with Hadoop -> rows can be input/output for map/reduce tasks

# HBase Data model

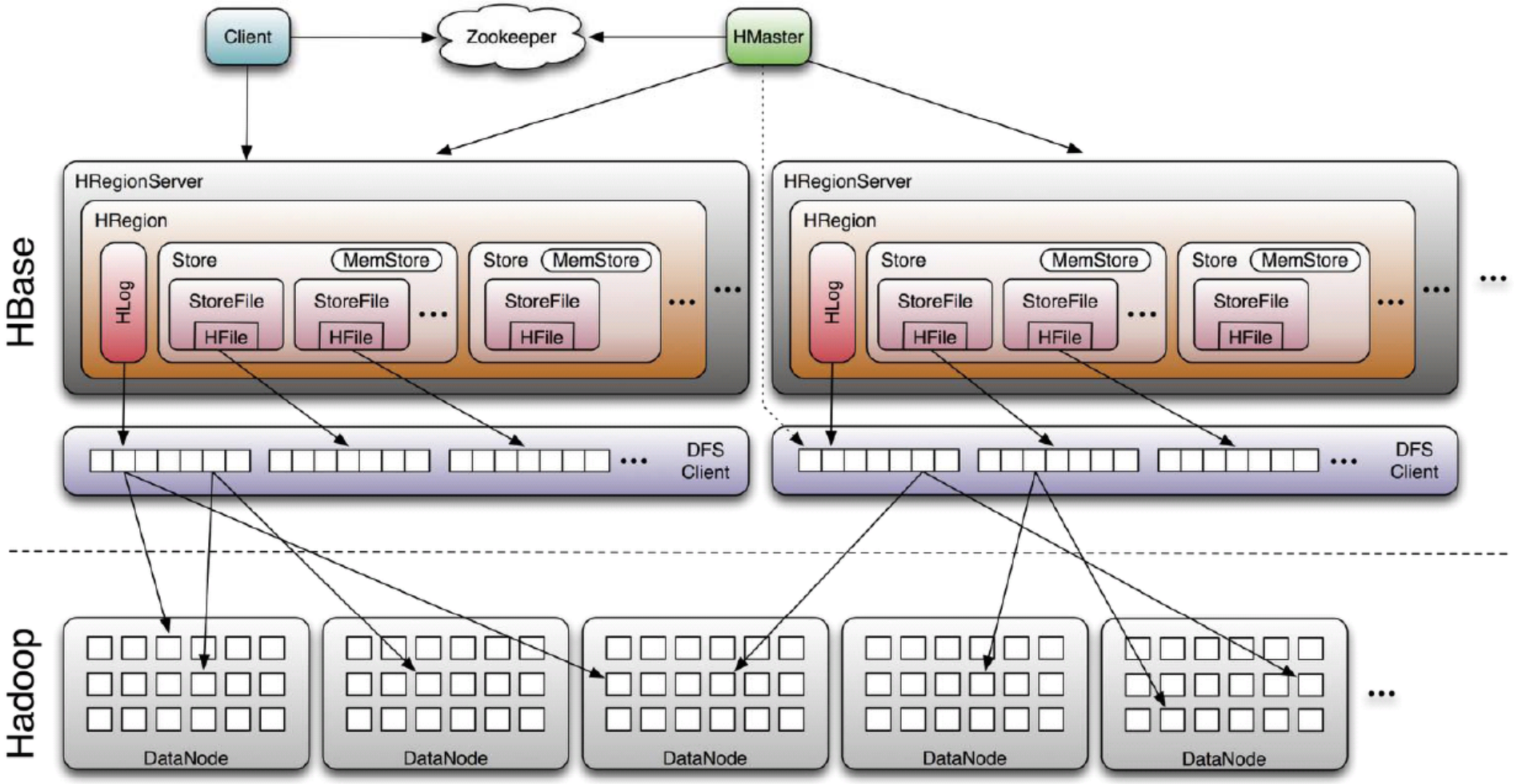


# HBase Data model (simplified)



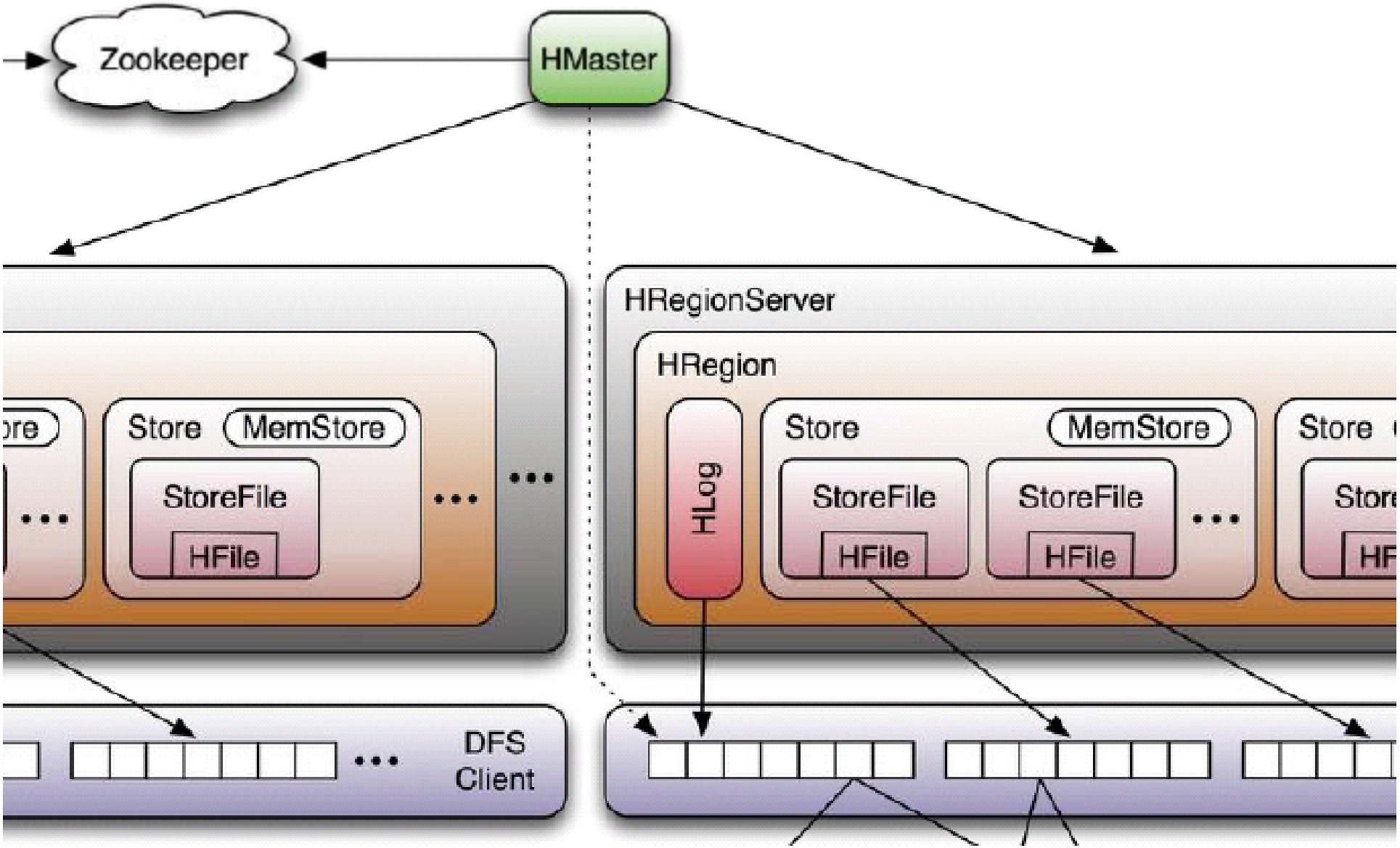

Regions partitioned on row key

# HBase System Architecture

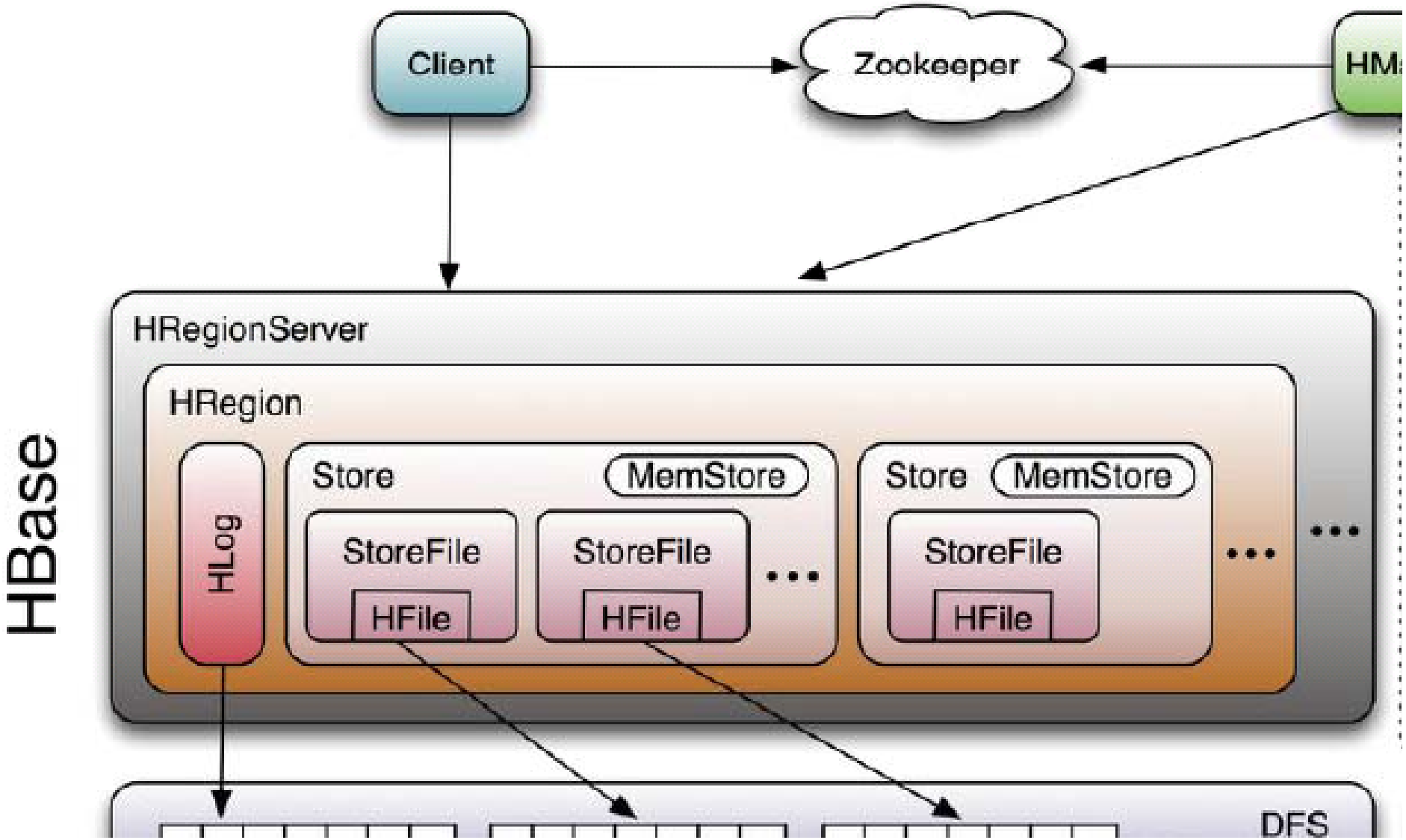



From http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html

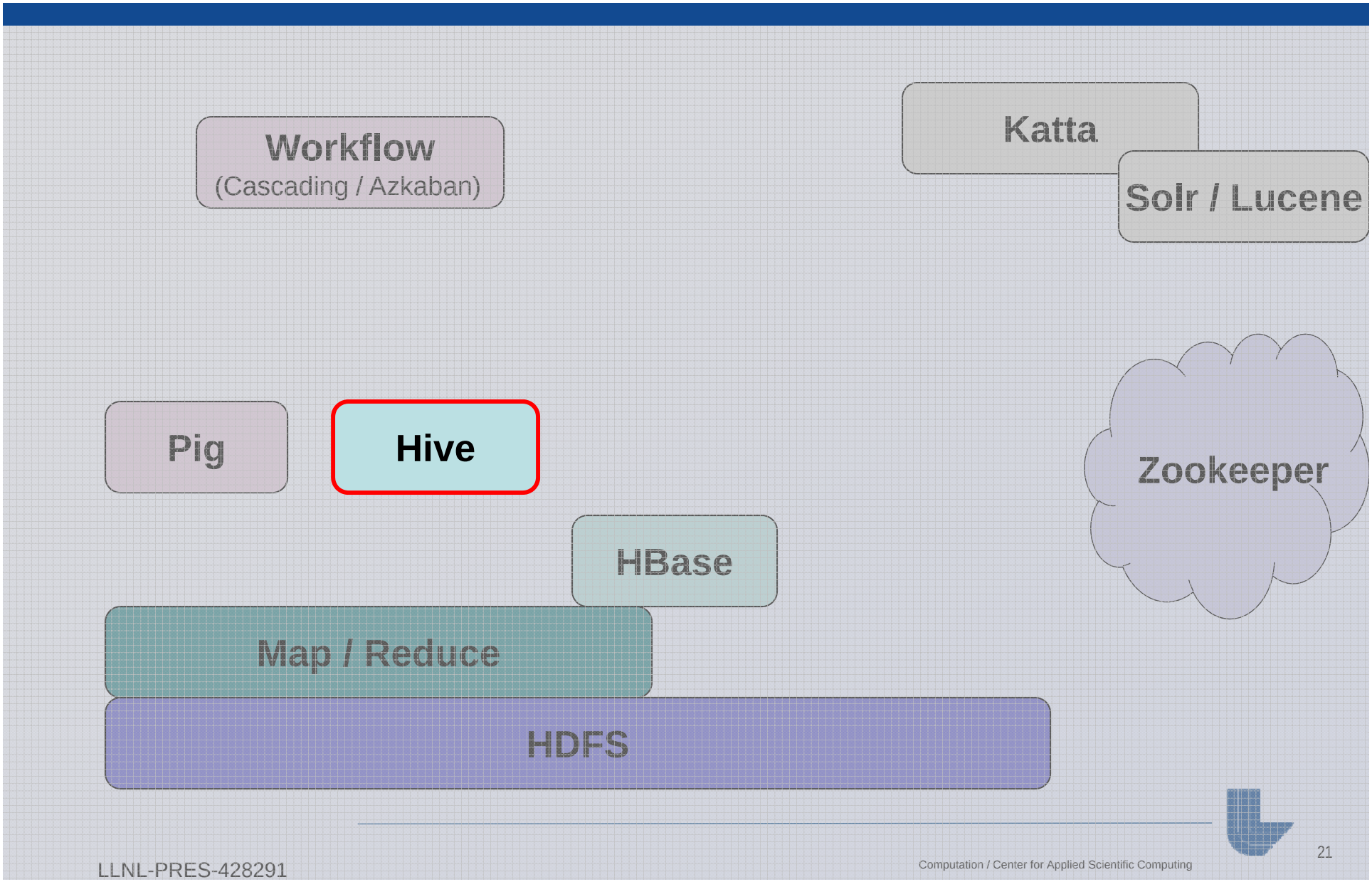
# HBase Master manages region servers



# Hbase Client directly access region servers for data



# The M/R stack of open source software – Hive

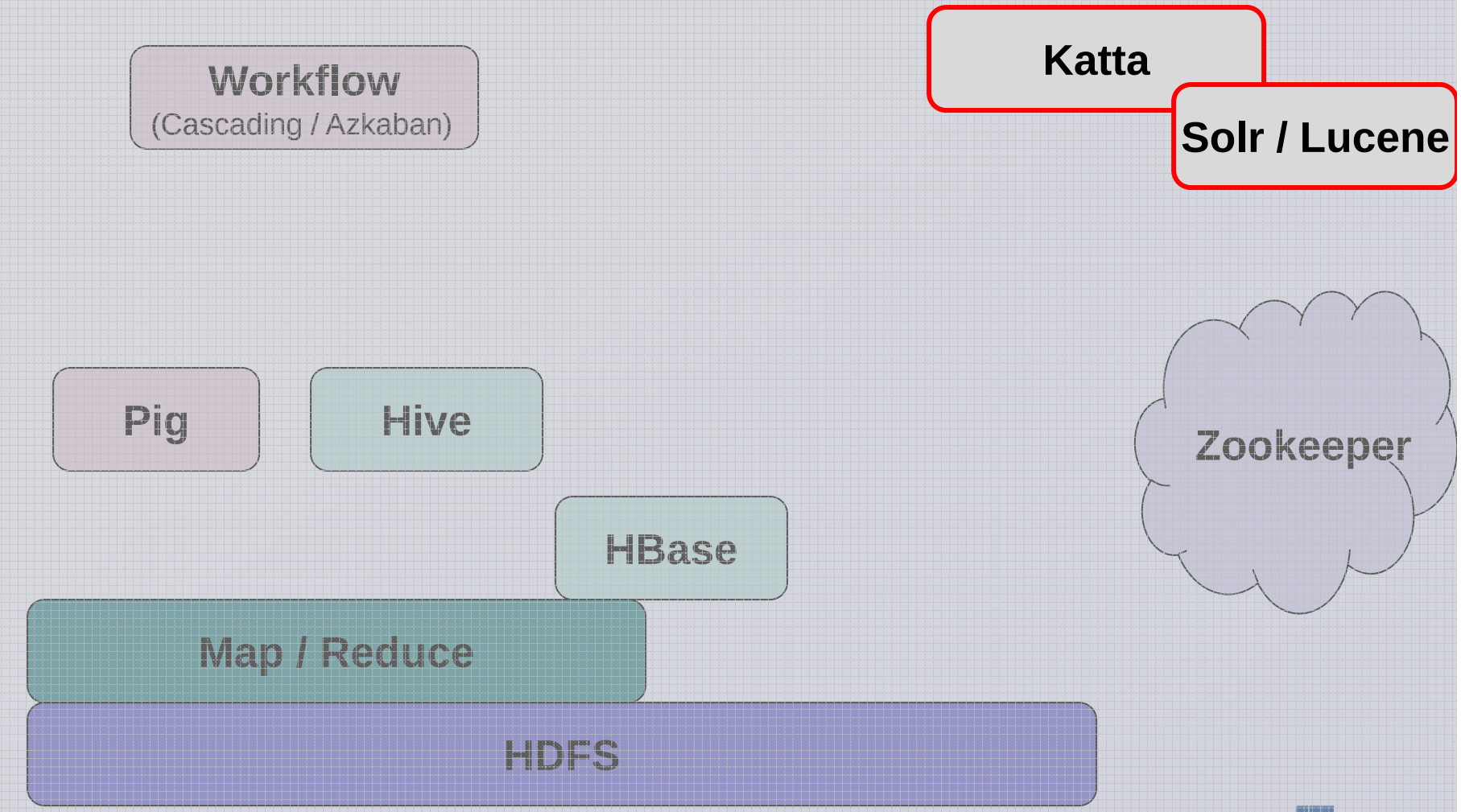


# Hive provides and SQL-like interface to data

- Components
  - Shell: SQL-like command line; Web;  JDBC
  - Driver:  API interface
  - Compiler: parse, plan, optimize
  - Execution Engine: DAG of stages (M/R, HDFS, or metadata)
  - Metastore: schema, location in HDFS, SerDe

http://www.cloudera.com/videos/introduction_to_hive

# The M/R stack of open source software – Solr / Katta



# Solr

- Faceted text search interface built on top of Lucene
- Built as a native web app – drops into any web server

# Faceted search is a foundational component for ad hoc document analysis

**Category**
‹ Any Category
**Electronics**
Audio & Video (82)
Computers & Add-Ons (5)
Accessories & Supplies (2)

**Brand**
‹ Any Brand
**Samsung**

**Seller**
**Any Seller**
ButterflyPhoto (23)
Amazon.com (22)
PRESTIGE CAMERA (16)
GENGLOBAL (16)
Crutchfield (15)
B&H Photo-Video (15)
Vanns (14)
› **See more...**

**Eco-friendly**
**All Products**
Energy Star (30)

**Price**
‹ Any Price
$ 1000 to $ 1999 GO

**Shipping Option**
**Any Shipping Option**
Can be shipped within one business day from Amazon.com (19)

**"televisions" › Electronics › Samsung › $1000-$1999**

**Related Searches:** lcd televisions, tv.

Showing 1 - 24 of 92 Results  « Previous | **Page: 1** 2 3

1.

**Samsung LNT5265F 52-inch 1080p LCD HDTV**
Buy new: ~~$2,999.99~~ Click to see price
11 Used & new from $1,574.06
In Stock
(183)

2.

**Samsung HLT5687SAX 56-inch**
Buy new: ~~$2,299.99~~ Click to see pri
5 Used & new from $1,190.05
In Stock
(151)

4.

**Samsung LNT4671F 46-inch 1080p 120Hz LCD HDTV**
Buy new: ~~$3,199.99~~ **$2,599.99**
6 Used & new from $1,609.95
In Stock

5.

**Samsung HLT6176S 61-Inch U**
Buy new: ~~$2,299.99~~ Click to see pri
4 Used & new from $1,488.00
In Stock

# Solr architecture

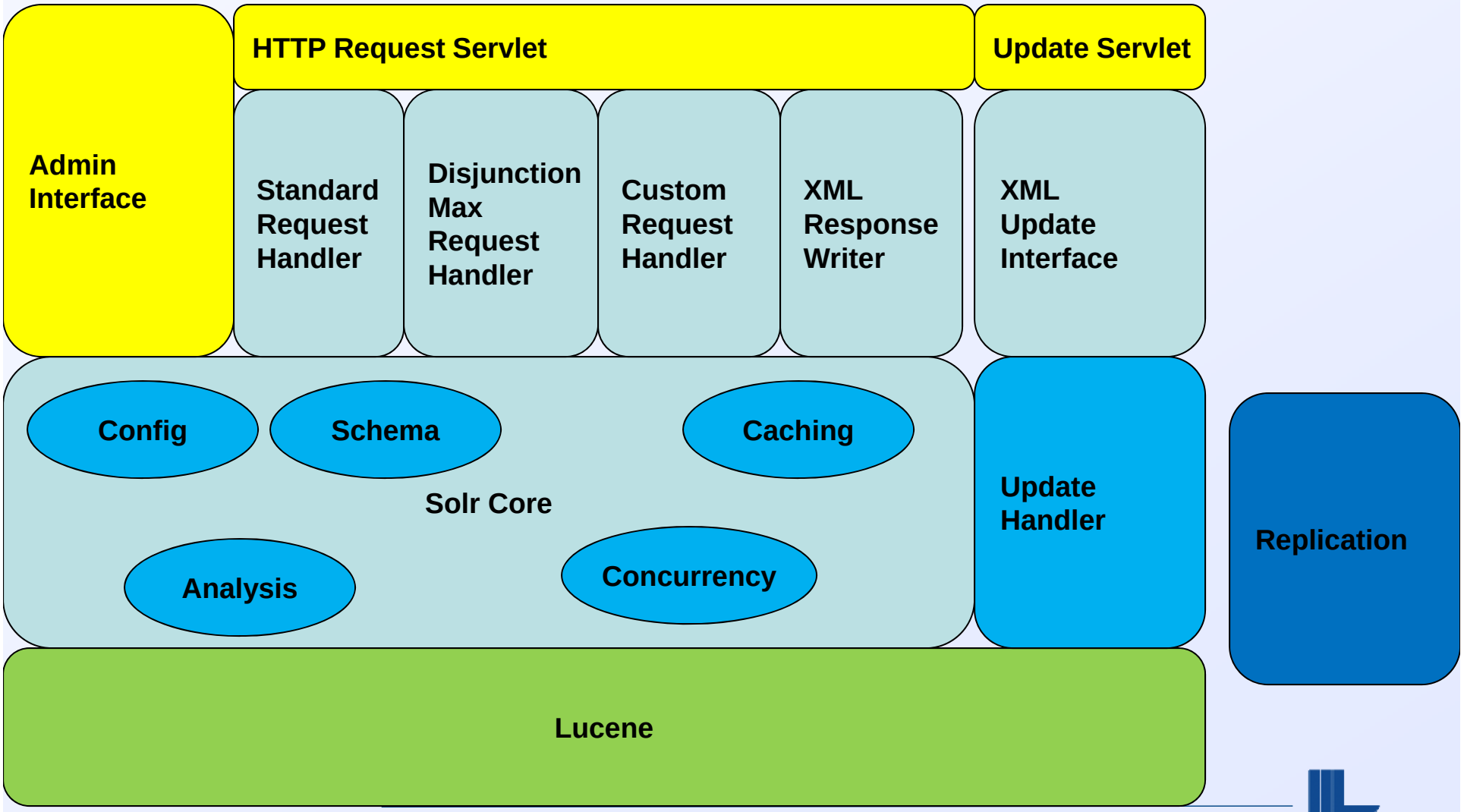
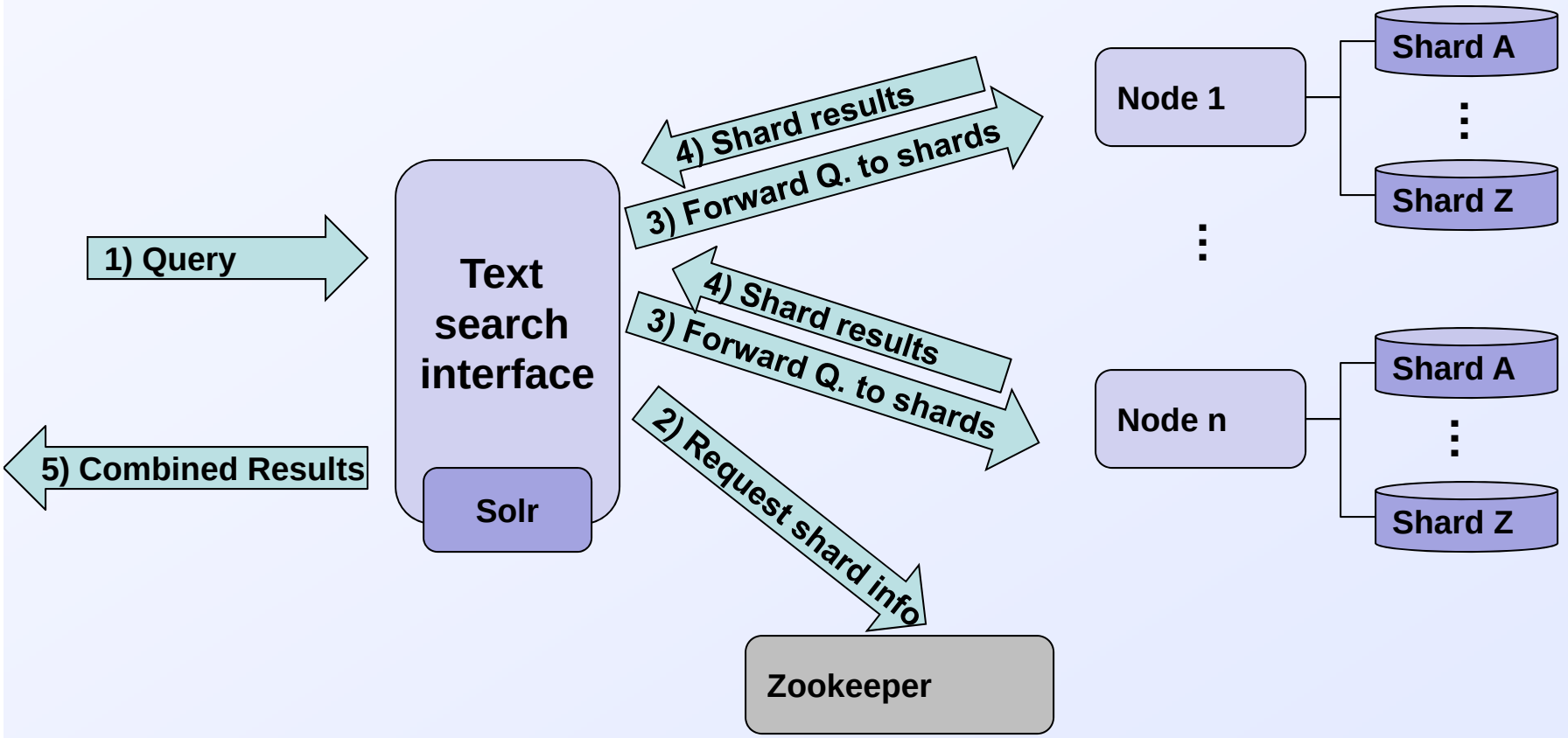



Diagram by Yonik Seeley

# Katta provides vertical and horizontal scalability



# Projects using Hadoop at LLNL

- Student projects
  - Bioinformatics [James Leek]
  - Continuous time LDA [Kurt Miller and Tina Elliasi-Rad]
- Advanced R&D projects
  - Network analytics
  - Keyword tagging and entity extraction
  - Faceted Search
- Research projects
  - READ LDRD
- Program deployments
  - BKMS

# Bioinformatics (student project)

- KPATH: produce DNA signatures for detection of pathogens
  - k-mer lexing: produce set of unique DNA sequences of length 15-60
    - Sliding window
  - Discover k-mers that are unique between bacteria and viruses

# K-mer parsing performance comparisons

- Lexing bacteria file 30 k-mer length [120 GB]
  - Optimized suffix tree [ C implementation]
    - on single node, 256 GB RAM, 16 processor system
    - 10.5 hours
  - Custom hadoop implementation
    - 85 nodes, 8 GB RAM, dual processor [old]
    - ~1 hour

# Unique K-mer grouping performance

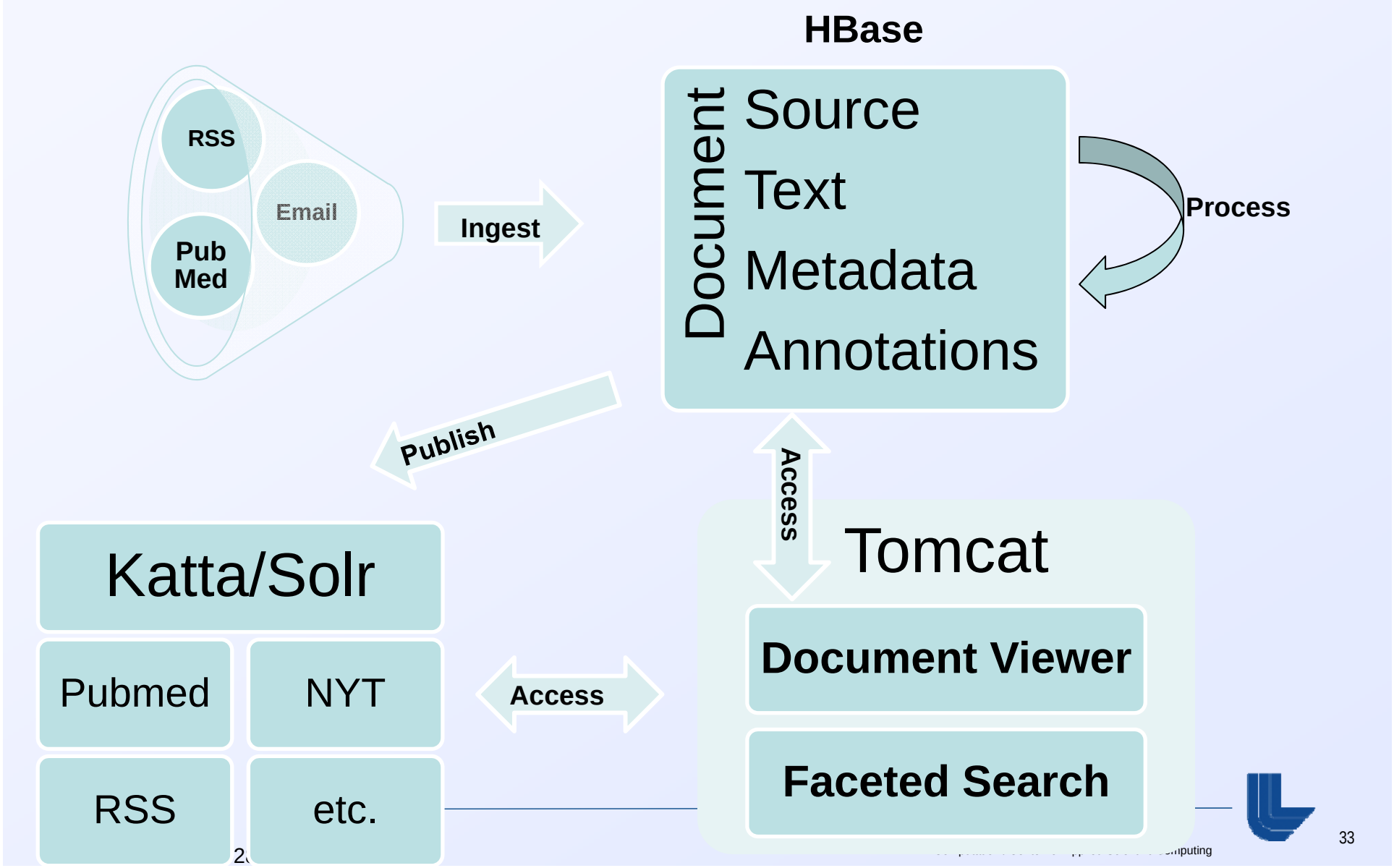
- Group unique k-mers of length 15 [13 GB data]
  - Pig implementation using outer joins [10 LOC]
    - More than 9 hours
  - Customer hadoop implementation:
    - 3 hours 26 minutes

# Network data

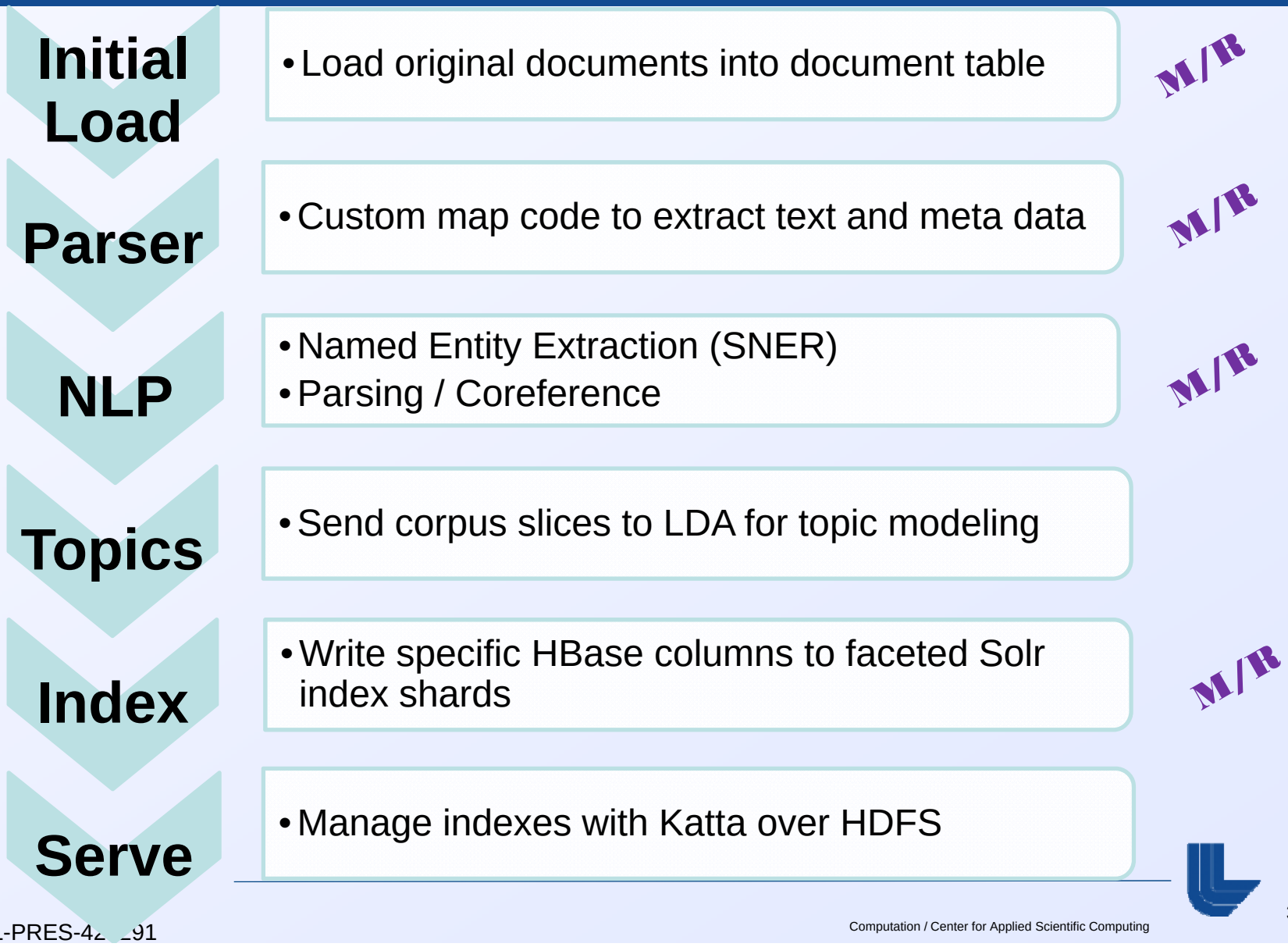
- HDFS provides storage layer for large repositories of network data
- Hive provides an SQL interface

- Performance on single query for 6 months of data:
  - Tuned Oracle DB: hours to days
  - Hive: minutes

# Hadoop-based document management architecture



# Example Data Flow



# Keyword tagging & Entity extraction

- Keyword tagging
  - Large dictionaries (100K terms)
  - Finite state machine to store dictionary

- Named Entity Recognition
  - Stanford NER
  - CRF model [People, Organizations, Location]

# Performance of Keyword tagging & Entity extraction

- 21M Pubmed entries + 1M news articles
  - 11M Pubmed abstracts
- 55K dictionary key phrases
- 6 node cluster [16 core, 96 GB RAM, 6 TB disk]

- Keyword Tagging
  - 8 minutes, 34 seconds
- Named Entity Annotation
  - 1 hr 58 minutes


**Query NYTSolr:**

Extracted | Source

person | organization | location

All | Hierarchical

| | | | |
|---|---|---|---|
| Abbas, Mahmoud | Abbas, Mahmoud (Prime Min) | Abt, Samuel | Akhta |
| Altman, Daniel | Anderson, Dave | Anderson, Michael P (Lt Col) | Anna |
| Araton, Harvey | Archibold, Randal C | Armstrong, Lance | Aske |
| Baker, Al | Banerjee, Neela | Barboza, David | Barri |
| Battista, Judy | Beers, Christopher | Bell, Jack | Belsc |
| Bennet, James | Berlusconi, Silvio (Prime Min) | Bernstein, Viv | Blair, |
| Bloomberg, Michael R | Bloomberg, Michael R (Mayor) | Bradsher, Keith | Brant |
| Bremer, L Paul III | Brick, Michael | Broder, John M | Brow |
| Bryant, Kobe | Bush, George W | Bush, George W (Pres) | Caldw |
| Carr, David | Chawla, Kalpana (Dr) | Cheney, Dick (Vice Pres) | Clare |
| Clark, Laurel Salton (Dr) | Clemens, Roger | Clinton, Bill | Coop |
| Curry, Jack | Curtis, Ben | Davis, Gray | Davis |
| Davis, James E (Councilman) | Dennehy, Patrick | Dewan, Shaila K | Dotsc |
| Dunlap, David W | Dunning, Jennifer | Elliott, Andrea | Elliott |
| Feder, Barnaby J | Feuer, Alan | Filkins, Dexter | Finley |
| Fuerbringer, Jonathan | Glater, Jonathan D | Greenhouse, Steven | Grime |
| Hakim, Danny | Hanley, Robert | Hart, Ariel | Healy |
| Hermoso, Rafael | Herszenhorn, David M | Hicks, Jonathan P | Hoge |


**Query NYTSolr:**

Extracted | Source

author | topics | taxonomy | online_section | print_section | series_name

All | Hierarchical

- Accidents and Safety
- Accounting and Accountants
- Acquired Immune Deficiency Syndrome
- Adve
- Agec
- Airlines and Airplanes
- Alcoholic Beverages
- All Star Games
- Appa
- Archaeology and Anthropology
- Architecture
- Area Planning and Renewal
- Arma Forces
- Atomic Weapons
- Automobiles
- Bankruptcies
- Bank
- Basketball
- Bicycles and Bicycling
- Biographical Information
- Biolo
- Bombs and Explosives
- Books and Literature
- British Open (Golf)
- Budg
- Cellular Telephones
- Child Abuse and Neglect
- Children and Youth
- City (
- Classical Music
- Coaches and Managers
- College Athletics
- Colle
- Computer Software
- Computers and the Internet
- Consumer Protection
- Cook
- Courts
- Credit
- Crime and Criminals
- Danc
- Decisions and Verdicts
- Demonstrations and Riots
- Diet and Nutrition
- Discr
- Drug Abuse and Traffic
- Drugs (Pharmaceuticals)
- Economic Conditions and Trends
- Edito
- Election Issues
- Elections
- Electric Light and Power
- Elect
- Ethics
- Executives and Management
- Finances
- Fines

## Extracted

| | |
|---|---|
| person | 29 |
| organization | 31 |
| location | 29 |

## Source

| | |
|---|---|
| author | 37 |
| topics | 202 |
| taxonomy | 254 |
| online_section | 52 |
| print_section | 40 |
| desk | 40 |

| Name | Not | Count |
|---|---|---|
| Business/Financial Desk | (X) | 31 |
| Editorial Desk | (X) | 4 |
| Metropolitan Desk | (X) | 3 |
| Foreign Desk | (X) | 2 |

| | |
|---|---|
| descriptor | 273 |

## Results

**iScore:** 0.5 Why?

**Id:** e0e92e1e7535ba6c2e8f8042aedf983b4a

**Solr:** NYTSolr

**Solr Score:** 6.085178

ABSTRACT:

A Philippine antigraft court has dealt a blow to efforts conglomerate San Miguel to regain control of the shar ago.

Coming after nearly 17 years of legal battles, the cou San Miguel chairman, Eduardo M. Cojuangco Jr., had a controlling stake in one of the country's largest banl the purchase and awarded ownership of the stake in

**iScore:** 0.5 Why?

**Id:** c8e609ae732796fab759447646a5eb57c

**Solr:** NYTSolr

**Solr Score:** 6.085178

# Faceted Search Indexing Performance

- Creating 1 Solr index on 1M news articles:
  - 8hrs 16 min
    - Map: 37 min
    - Reduce: 8 hrs 14 min
- Creating 50 Solr indexes on 1M news articles:
  - 55 min
    - Map: 7 min
    - Reduce: 54 min

# Open-sourced products and others in the open source pipeline

- iScore
  - Content-based personalization
  - [pre-Hadoop]
- Reconcile
  - Coreference resolution software built on open source tools [with Cornell and U. Utah]
  - Additional adaptation to Hadoop
- Dunk
  - An elegant java annotation system that allows you to have the fields of a java object serialized (deserialized) to (from) an HBase table
  - Simplifies queries, object construction, and map/reduce formulation

# Questions?

## Disclaimer