CSC 555 and DSC 333
Mining Big Data
Lecture 9

Alexander Rasin

College of CDM, DePaul University

November 9th, 2021

Tonight

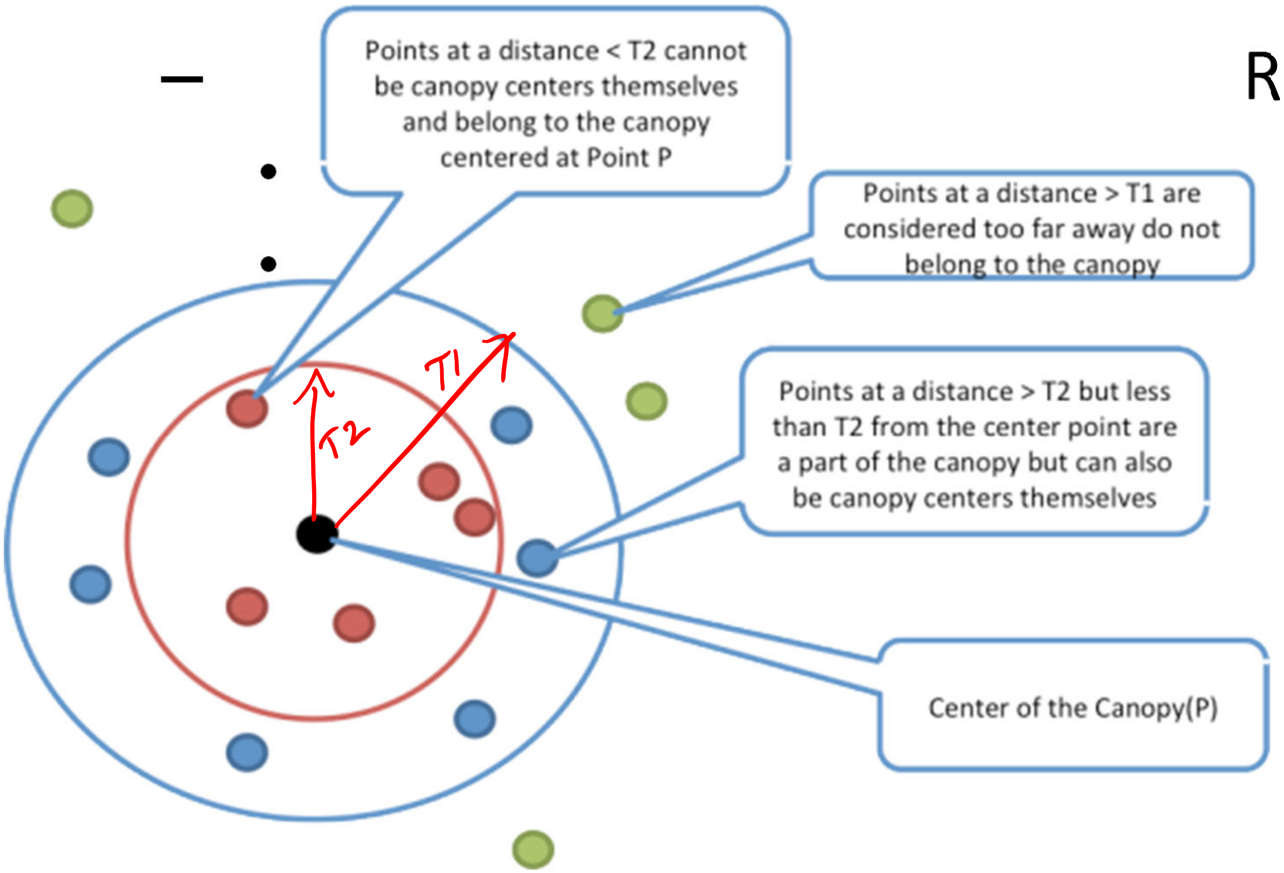
- Clustering
- Document matching

Map-side join

- Running Spark
- Recommender systems

Canopy (Pre-)Clustering

- Used for pre-processing
  - Initialize centroids
  -



Repeat

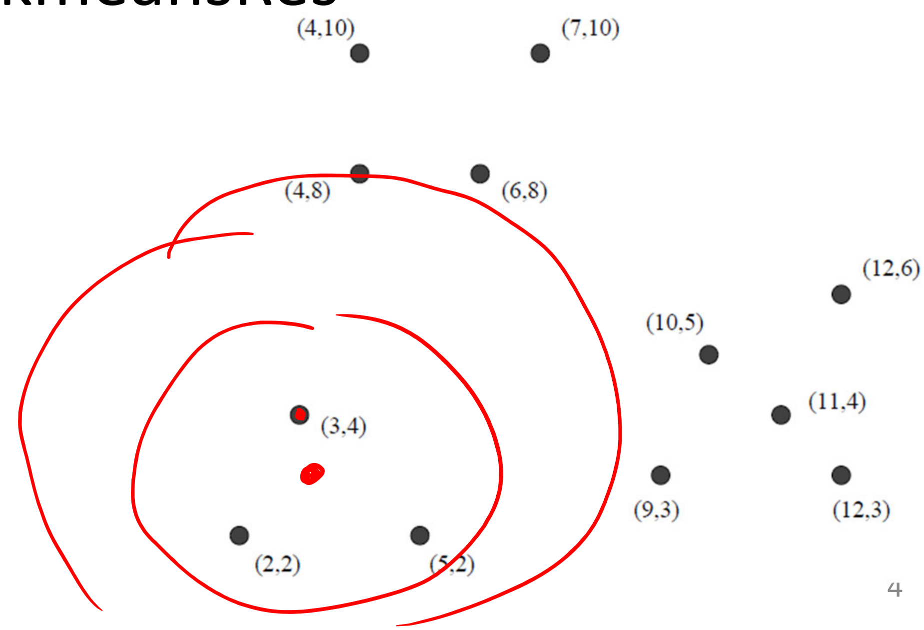
Pick random center

Build canopy

Remove points

K-Means Clustering with Mahout

- Create a simple input file (12 points)
- $MAHOUT_HOME/bin/mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --maxIter 8 --numClusters 3 --t1 5 --t2 3 --input testdata --output kmeansRes

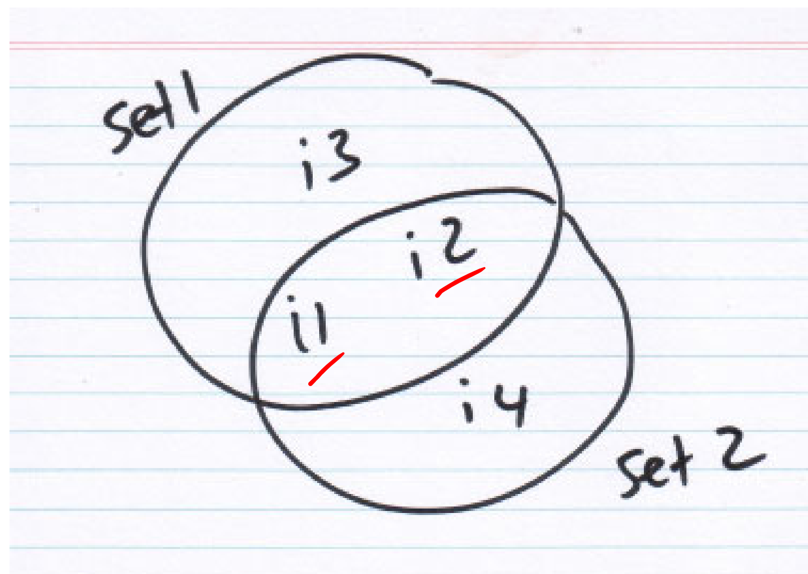


Similarity of Sets

- How do you compare two sets?

$\frac{2}{4}$

- Items represented by sets
  - Documents
  - Homeworks
  - Fingerprints
  - SQL Queries



- Overlap = “Similarity”

Jaccard Similarity Measure

- Clustering/recommender engines
- Find buyers with similar taste
  - Collaborative/content filtering
- Find movie-renters with similar taste
  - Netflix, Blockbuster
  - Ratings are 1-5, not boolean
    - Bag distance (minimum for intersection, sum for union)
    - {a,a,a,b} <> {a,a,b,b,c} = 1/3

Shingling

- A mechanism to represent documents
- Pick value k
- Generate k-shingles to represent the document
  - Document = abcdabd
  - 2-shingles = {ab, bc, cd, da, bd}
- Compute similarity
- White space? (‘ ‘, ‘\n’, …)

Shingle Size

- What if we pick k=1?
- K large enough to keep shingle appearance probability low
- How many characters do you expect?
  - 30^2 = 900
  - 30^3 = 27K
  - 30^4 = 810K
  - 30^5 = ~24M combinations

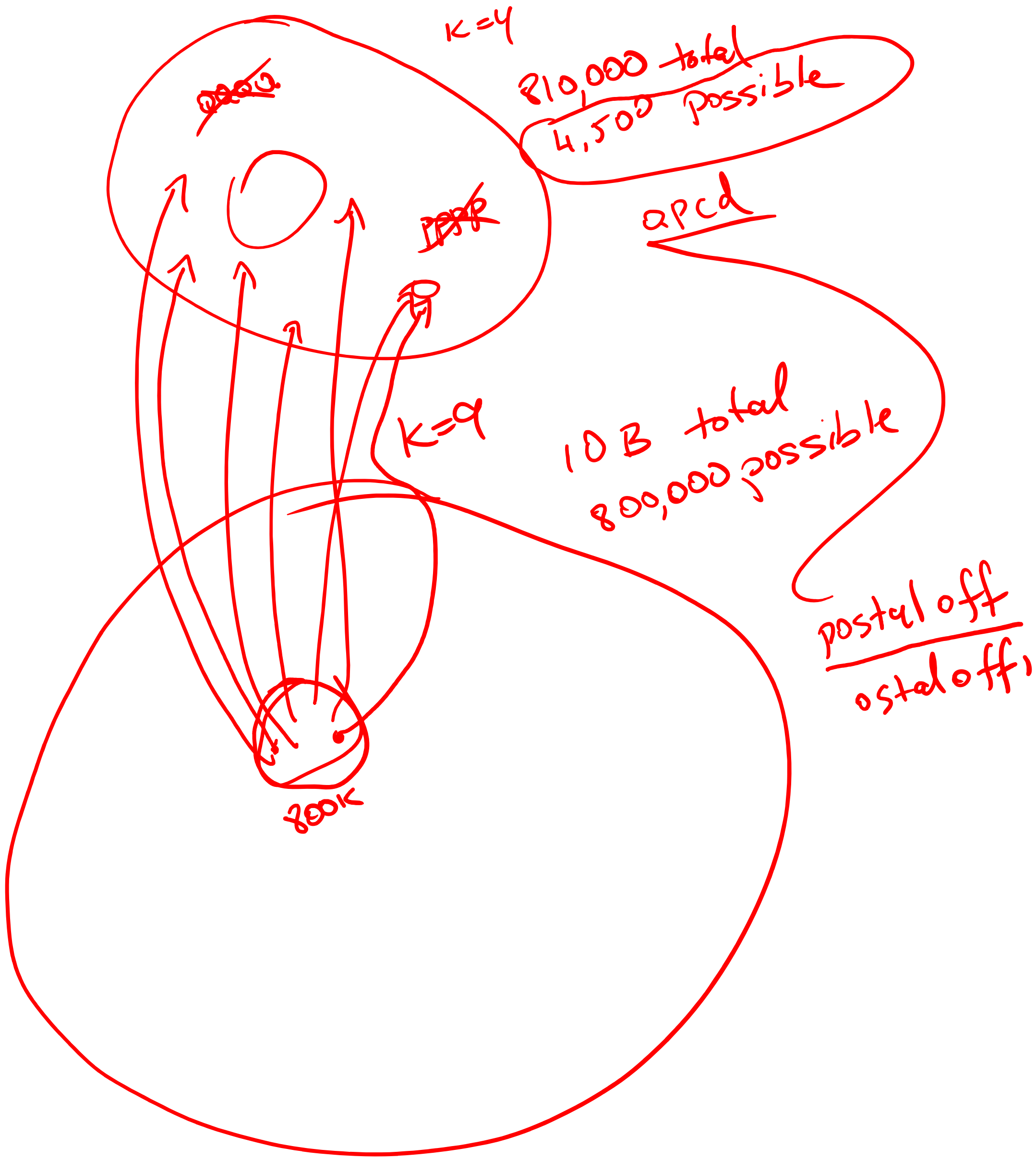
Letter Distribution

- Letter distribution is not uniform
  - Scrabble values
- 810K combinations

  ZQZP
- Common 4-letter combinations are a much smaller subset
  - Others are unusual or impossible

String Overhead

- Comparing strings is expensive
  - For k=9, abcdefghi and abcdefghz
- One of the many uses of hashing
  - Similar to compression (lossy)
  - Represent each shingle by a code (hash)
- How much space does the hash require?
  - k=9, 30^9 different potential shingles
- k=9 hashed to 4 bytes better than k=4
  - Effective space is much smaller



K=4
810,000 total
4,500 possible
apcd
K=9
10B total
800,000 possible
postal off
ostaloff,
800K

Shingle Cost

abcd
bcde
cdef

- k=9 -> 9X document size
  - Hash to 4 bytes, still 4X
  - Does not fit in memory
- Create signatures
  - Again, hashing
  - Lossy compression
  - Similarity-preserving

Matrix Representation

- Each document = binary vector
  - – # elements
  - – # of documents
- Jaccard measure
- Sparse matrix

postal off

zzqt

$\{a, d\}$ $\{c\}$ $\{b, d, e\}$

abcd

postab

Element	S_1	S_2	S_3	S_4
a	1	0	0	1
b	0	0	1	0
c	0	1	0	1
d	1	0	1	1
e	0	0	1	0

Figure 3.2: A matrix representing four sets

Minhashing

- Select a permutation of the matrix rows
- The first occurrence of 1 in the vector
  - $h(S_1) = a$
  - $h(S_2) = c$
  - $h(S_3) = b$
  - $h(S_4) = a$
- Similar to the Jaccard measure

Element	S_1	S_2	S_3	S_4
b	0	0	1	0
e	0	0	1	0
a	1	0	0	1
d	1	0	1	1
c	0	1	0	1

Figure 3.3: A permutation of the rows of Fig. 3.2

Minhash Signatures

5 rows

5 rows

- Permute and build a Minhash signature several times
- No need to build the permutation

Row	S_1	S_2	S_3	S_4	$x + 1 \mod 5$	$3x + 1 \mod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

Figure 3.4: Hash functions computed for the matrix of Fig. 3.2

- Final result (2)

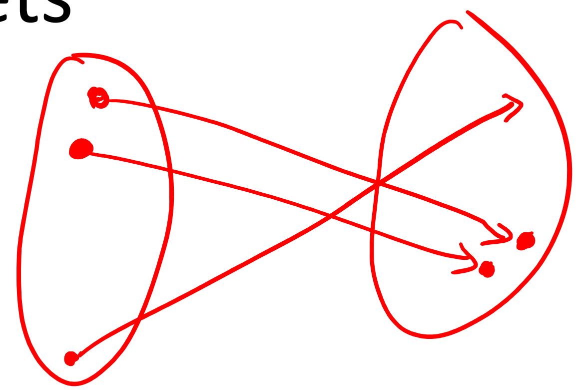
	S_1	S_2	S_3	S_4
h_1	1	3	0	1
h_2	0	2	0	0

Comparing Document Pairs

- The curse of the N^2
  - 100,000 documents
  - Enough memory to store hash/signatures
  - ~100,000 x 100,000 pairs = ~10,000,000,000 comparisons
- This can take a long time
  - 1000 comparisons/second
  - 115 days
- (embarassingly) Parallelizable

Locality-Sensitive Hashing

- Assign documents into “buckets”
- 100K => 50 buckets
  - 100K x 100K
  - 50 * 2,000 * 2,000 = 200,000K
  - (vs 10,000,000K)
  - 115 days => 2.3 days
- Split the signature matrix into “bands”
- Hash each band



Combining the Techniques

- Construct set of k-shingles (for some k)
  - Optionally hash shingles to n-bit values
- Arrange the documents by shingle value
- Pick a length for the minhash signatures
- Select a threshold t (false neg. vs speed)
- Construct candidate pairs by applying LSH
- Find the candidate matches
  - Optionally, look at the actual matched documents

Bloom Filter

- Simple hashed approximation
- Find match (with false positives)
- Example
  - 10M URLs
  - 8bits/URL (~10MB)
    - ~2% false positive rate
  - 10bits/URL (~12.5MB)
    - ~0.8% false positive

0 1 1 1 1 1 0 0

depaul.edu

— URL4

h1(url1)

h2(url1)

7:30

A Break

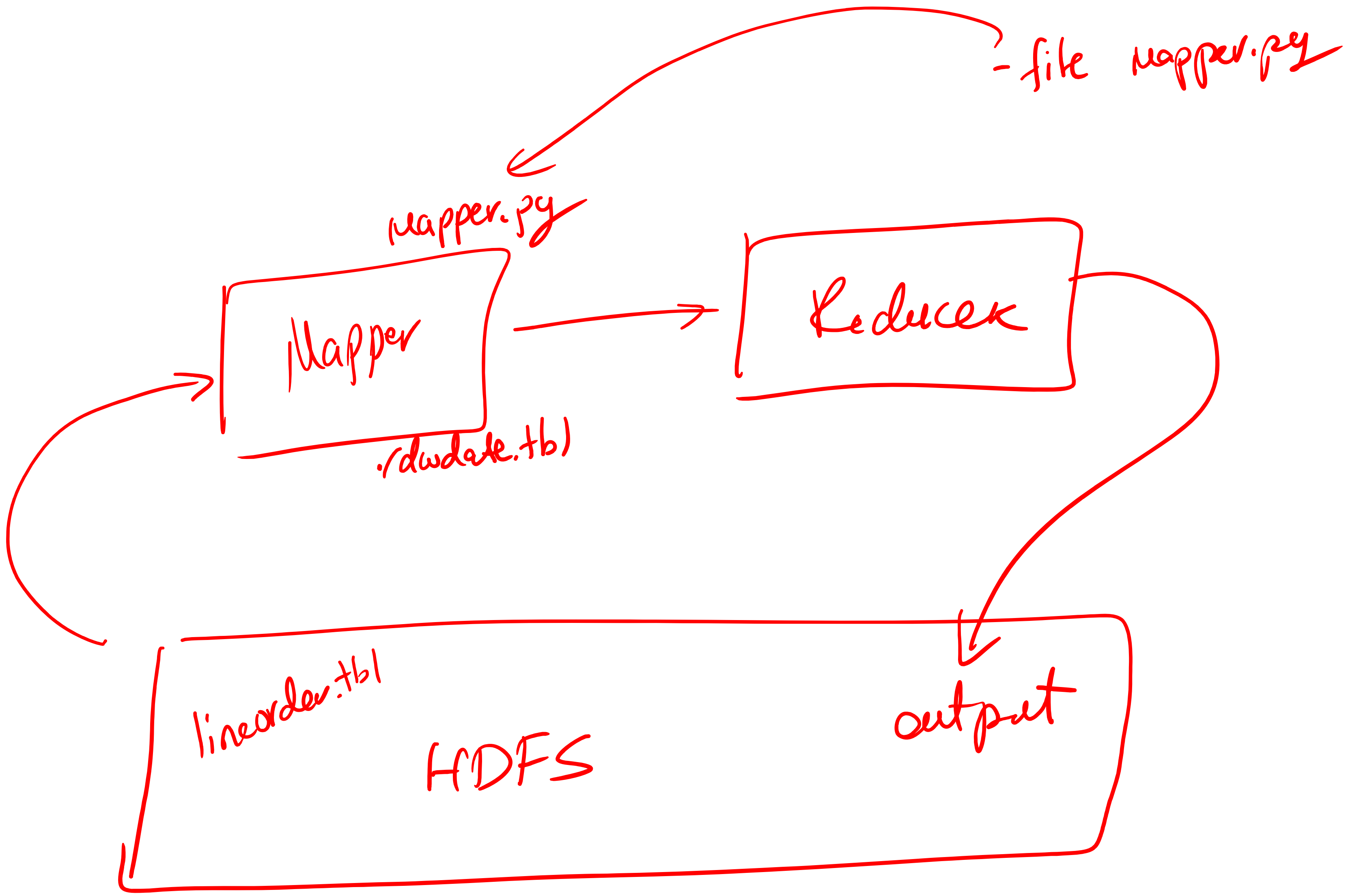
Map-join implementation

SELECT lo_quantity, AVG(lo_revenue)

FROM lineorder, dwdate

WHERE lo_orderdate = d_datekey AND d_year = 1994
AND lo_discount BETWEEN 6 AND 8

GROUP BY lo_quantity;


-file mapper.py
mapper.py
Mapper
./dwdate.tbl
Reducer
lineorder.tbl
HDFS
output

Spark Principles

file = 8 blocks × 128 MB
in HDFS/MR

in spark file = 100M lines × line

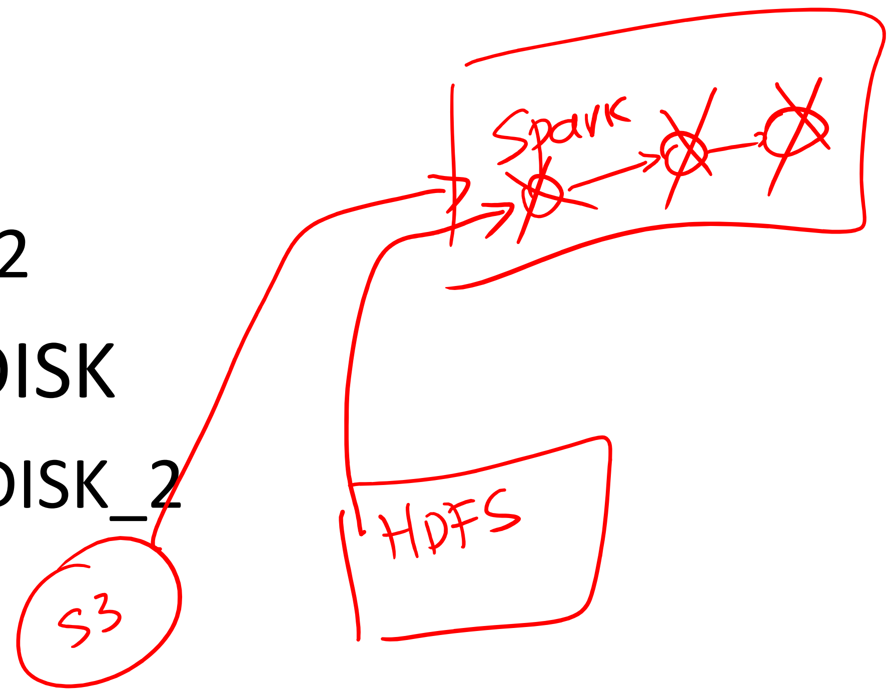
- Distributed datasets
  - data = [1, 2, 3, 4, 5]
  - distData = sc.parallelize(data)
  - distData.reduce(lambda a, b: a + b)
- distFile = sc.textFile("data.txt")
  - hdfs://... , s3n://
  - lengths = distFile.map(lambda s: len(s))
  - totalLength = lengths.reduce(lambda a, b: a + b)

Spark Storage

- As close to data as possible (HDFS nodes?)
- Uses local disk to store data
  - Recommended 4-8 disks per node w/out RAID
- Recommend allocating up to 75% of RAM
- When data in RAM, performance network-bound

RDD Persistence

- MEMORY_ONLY
  - MEMORY_ONLY_2
- MEMORY_AND_DISK
  - MEMORY_AND_DISK_2
- DISK_ONLY
- Unpersist



Spark Example

```
# Read file from HDFS
text_file = sc.textFile("hdfs://ip-172-31-29-219.us-west-1.compute.internal/data/README.md")
lengths = text_file.map(lambda s: len(s))
print text_file.take(100)

lengths.foreach(myprint)
print lengths.take(100)
totalLength = lengths.reduce(lambda a, b: a + b)
```

Spark Example

```
# Read file from HDFS
text_file = sc.textFile("hdfs://ec2-54-67-64-123.us-west-1.compute.amazonaws.com/data.txt")
counts = text_file.flatMap
          (lambda line: line.split(" ")).map(
           lambda word: (word, 1)).
           reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://ip-172-31-29-219.us-west-1.compute.internal/data/output")
```

SUM

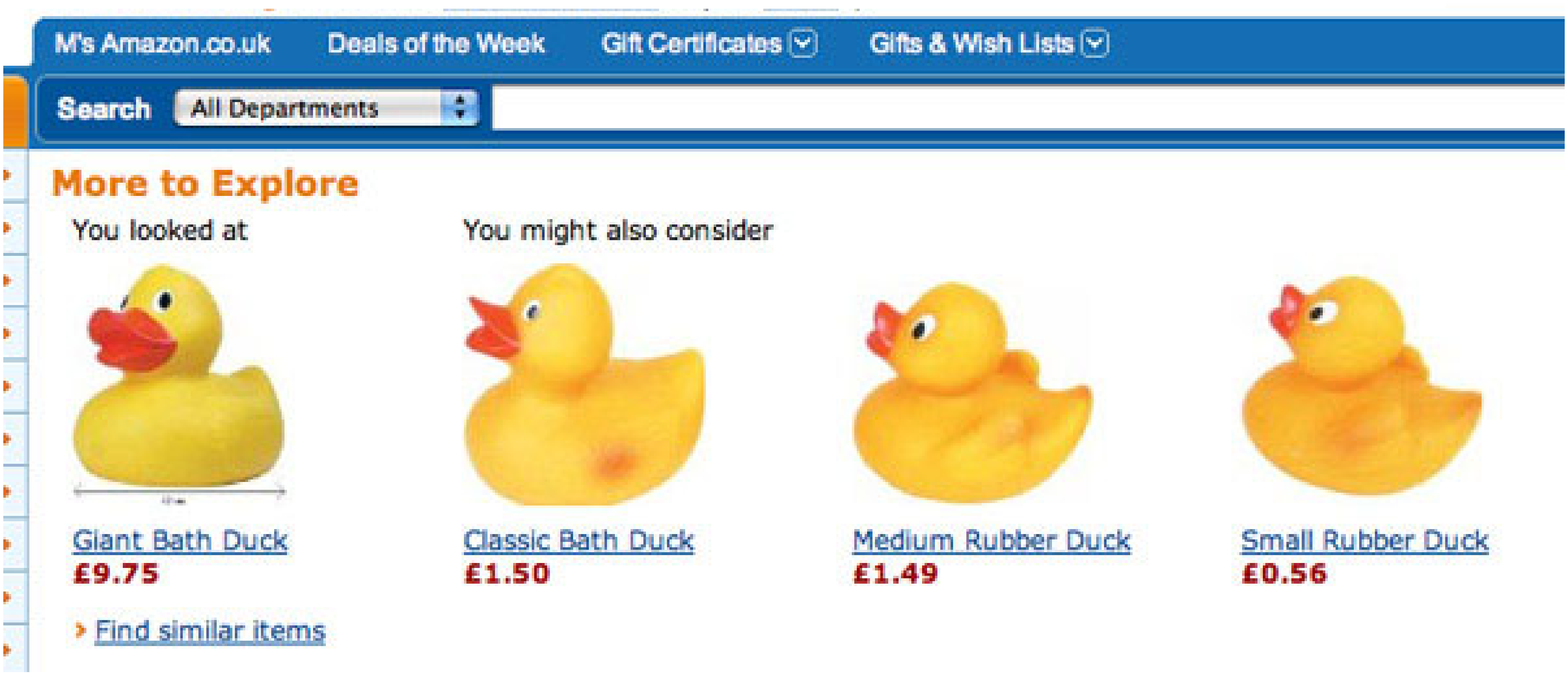
Spark Example

```
import random
def sample(p):
    x, y = random.random(), random.random()
    return 1 if x*x + y*y < 1 else 0

count = sc.parallelize(xrange(0, 500000)).map(sample) \
            .reduce(lambda a, b: a + b)
print "Pi is roughly %f" % (4.0 * count / 500000)
```

Recommender Systems

- Recommend
  - Movies/Purchases/News
- Content-based systems
  - Analyze user history
  - Find similar items
- Collaborative filtering
  - Similarity between users
  - Users like you also liked…


M's Amazon.co.uk
Deals of the Week
Gift Certificates
Gifts & Wish Lists
Search
All Departments
More to Explore
You looked at
You might also consider
Giant Bath Duck
£9.75
Find similar items
Classic Bath Duck
£1.50
Medium Rubber Duck
£1.49
Small Rubber Duck
£0.56

Coyote Urine Lure 16 oz.

Deerbusters

No customer reviews yet. Be the first.

List Price: ~~$19.95~~

Price: $15.95

You Save: $4.00 (20%)

In Stock.

Ships from and sold by **MasterGardening**.

Customers Who Bought This Item Also Bought

Shake Away 9002020 20oz Cat Repellent Coyote / Fox Urine

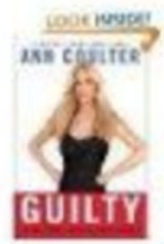
(14)

$14.99

Coyote Urine Lure-32 oz

(4)

$29.95



Guilty: Liberal "Victims" and Their Assault on Ame... by Ann Coulter

(369)

$10.88


Amazon.com: Why is this recommended for you?
https://www.amazon.com/gp/yourstore/rate-this-asin/ref=pd_ys_...
amazon.com
Help | Close window
Recommended for You
Saw: The Final Chapter
Our Price: $3.99
See all buying options
Rate this item
I own it
Not interested
Because you purchased...
Polar Express (Video On Demand)
This was a gift
Don't use for recommendations
Help | Close window


amazon.com
Hello. Sign in to get personalized recommendations. New customer? Start here.
Your Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards
Shop All Departments
Search Sports & Outdoors
Sports & Outdoors
Athletic & Outdoor Clothing
Bikes & Scooters
Exercise & Fitness
Large Crowbar
Other products by Emergency Disaster Systems, Inc.
No customer reviews yet. Be the first. | More about this product
Price: $12.00
In Stock.
Ships from and sold by Emergency Disaster Systems, Inc..
Save up to 70%
Up to 70% Savings on Thousands of Products
Find great bargains on thousands of products in Sports & Outdoors orders. Shop now.
See larger image
Share your own customer images
Frequently Bought Together
Customers buy this item with The Zombie Survival Guide: Complete Protection from the Living Dead by Max Brooks
Price For Both: $22.04
Add both to Cart
Add both to Wish List
These items are shipped from and sold by different sellers. Show details

Utility Matrix

- Users/ratings
- Very sparse

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Populating the Utility Matrix

- Determine the (relevant) features
- Populate the values
  - User purchase
  - User like/dislike
  - User rating

Collaborative Filtering

- Jaccard measure loses information

$\frac{1}{5}$

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- A<->B
  - Jaccard similarity of 1/5 (distance of 4/5)
  - Yet they agree on HP1 (the only common movie)
- A<->C
  - Jaccard distance of 1/2

Rounding the Data

- Replace
  - 1, 2 => No rating
  - ~~3, 4, 5 => 1~~

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	1			1			
B	1	1	1				
C					1	1	
D		1					1

- Jaccard
  - A to B distance => 3/4
  - A to C distance => 1

Normalizing Ratings

- Subtract the average from each value
  - How "different" is the rating

$\frac{14}{3} = 4\frac{2}{3}$

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

Clustering Users/Items

- Utility matrix is very sparse
  - Unlikely to find many matches
  - Cluster to unite attributes

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

	HP	TW	SW
A	4	5	1
B	4.67		
C		2	4.5
D	3		3

Clustering Users/Items

- Hierarchical clustering
- Revised matrix is denser
- Can also cluster users in the same manner
- Can repeat the process

	HP	TW	SW
A	4	5	1
B	4.67		
C		2	4.5
D	3		3

Collaborative Filtering

- MovieLens data
  - (User, Movie, Rating, Date)
  - Predict/recommend movies
- Netflix challenge
  - 480,000 users
  - 18,000 movies
  - 100M ratings
  - Minimize RMSE (2.8M testing set)
    - Netflix's CineMatch scored 0.9514

Netflix Challenge

- Data had been removed
- Removing user info does not “anonymize”
  - Can reverse-engineer users
  - With 8 movie ratings and a up to 14 day error dates => 99% can be identified
  - Two ratings with 3 day error => 68% identified
  - 6 movies outside of top 500 (without dates) => 84% accuracy
- Can mine IMDB for data

Matrix Decomposition Example

Users X Items

	M1	M2	M3	M4
A	5.00	5.00	2.00	-?
B	2.00	-?	3.00	5.00
C	-	5.00	-	3.00
4	3.00	-	-	5.00

Users X Features

	F1	F2	F3
A	1.12	1.49	0.48
B	1.31	-0.52	0.59
C	1.13	0.67	-0.52
D	1.39	0.05	0.45

Features X Items

	M1	M2	M3	M4
F1	1.81	2.66	1.73	3.16
F2	1.62	1.71	-0.23	-0.24
F3	0.74	-1.08	0.78	0.90

4.78	5.01	1.97	3.61
1.97	1.96	2.85	4.80
2.75	4.71	1.40	2.94
2.93	3.30	2.74	4.79

Root Mean Squared Error

- Evaluate error between estimator and actual values
  - Vectors:

$$\theta_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix} \quad \text{and} \quad \theta_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}.$$

$$\sqrt{\frac{\sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2}{n}}$$

$\boxed{0.22^2}$ $\boxed{0.01^2}$

Next Time:

- Larger Hadoop Ecosystem Overview
- Web advertising
- Mining Social Graphs