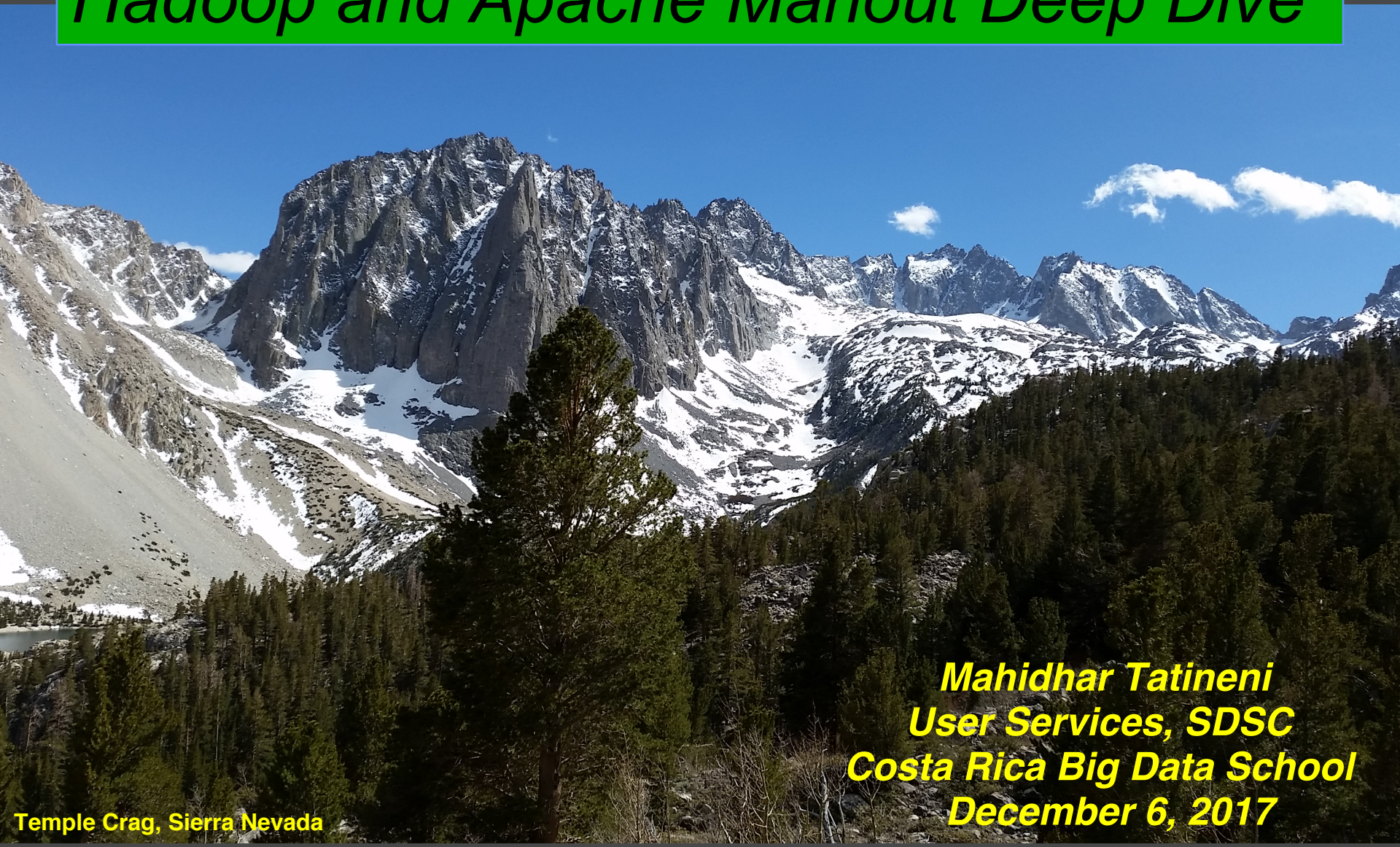
Hadoop and Apache Mahout Deep Dive



Mahidhar Tatineni
User Services, SDSC

Costa Rica Big Data School
December 6, 2017

Temple Crag, Sierra Nevada

Overview

- **Hadoop configuration files**
  - core-site.xml
  - hdfs-site.xml
  - yarn-site.xml
- **YARN architecture**
- **HDFS2 architecture**
- **HDFS admin commands, fsck**
- **Apache Mahout Algorithms and Examples**
  - Summaries of algorithms
  - Classification Example
  - Recommendation Example

Hadoop Typical Configuration Files

```
capacity-scheduler.xml
configuration.xsl
container-executor.cfg
core-site.xml
hadoop-env.cmd
hadoop-env.sh
hadoop-metrics.properties
hadoop-metrics2.properties
hadoop-policy.xml
hdfs-site.xml
httpfs-env.sh
httpfs-log4j.properties
httpfs-signature.secret
httpfs-site.xml
kms-acls.xml
kms-env.sh
kms-log4j.properties
kms-site.xml
log4j.properties
mapred-env.cmd
mapred-env.sh
mapred-queues.xml.template
mapred-site.xml
mapred-site.xml.template
masters
myhadoop.conf
slaves
ssl-client.xml.example
ssl-server.xml.example
yarn-env.cmd
yarn-env.sh
yarn-site.xml
```

- *Don't worry - usually not the job of end users to change these parameters!*
- *Typically commercial hadoop installations have GUI for changing values.*

core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/scratch/mahidhar/13124736/tmp</value>
  <description>A base for other temporary directories.</description>
</property>

<property>
  <name>fs.defaultFS</name>
  <value>hdfs://comet-10-02.ibnet:54310</value>
</property>

</configuration>
```

core-site.xml parameters

Parameter	Default	Description/Options
hadoop.security.authentication	simple	simple, kerberos
io.file.buffer.size	4096	multiple of page size
hadoop.tmp.dir	/tmp/hadoop-${user.name}	A base for other temporary directories.
io.seqfile.local.dir	${hadoop.tmp.dir}/io/local	Local directory for storing intermediate data files
fs.s3.block.size	67108864	Block size to use when writing files to S3.
hadoop.http.authentication.type	simple	simple/kerberos Auth for Oozie HTTP endpoint
nfs.exports.allowed.hosts	*rw	Remote mount of filesystem

Lot more options. See:
https://hadoop.apache.org/docs/r2.8.0/hadoop-project-dist/hadoop-common/core-default.xml

hdfs-site.xml

```
<configuration>

  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/scratch/mahidhar/13124736/namenode_data</value>
    <description>Determines where on the local filesystem the DFS name node
      should store the name table.  If this is a comma-delimited list
      of directories then the name table is replicated in all of the
      directories, for redundancy. </description>
    <final>true</final>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/scratch/mahidhar/13124736/hdfs_data</value>
    <description>Determines where on the local filesystem an DFS data node
       should store its blocks.  If this is a comma-delimited
       list of directories, then data will be stored in all named
       directories, typically on different devices.
       Directories that do not exist are ignored.
    </description>
    <final>true</final>
  </property>

  <property>
   <name>dfs.namenode.secondary.http-address</name>
   <value>comet-10-02.ibnet:50090</value>
   <description>The secondary namenode http server address and
       port.</description>
   <final>true</final>
  </property>

</configuration>
```

hdfs-site.xml parameters

Parameter	Default	Description/Options
dfs.default.chunk.view.size	32768	Number of bytes to view for a file on the browser
dfs.namenode.name.dir	file://${hadoop.tmp.dir}/dfs/name	Determines where on the local filesystem the DFS name node should store the name table
dfs.namenode.fs-limits.min-block-size	1048576	Minimum block size in bytes
dfs.replication	3	Default block replication
dfs.heartbeat.interval	3	Datanode heartbeat interval in seconds
dfs.client-write-packet-size	65536	Packet size for clients to write
dfs.encrypt.data.transfer	false	Should block data that is read/written from/to HDFS be encrypted?

Lot more options. See:
https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml

yarn-site.xml

```
<configuration>

<!-- Site specific YARN configuration properties -->
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>comet-10-02.ibnet</value>
    <description>The hostname of the RM.</description>
    <final>true</final>
  </property>

  <property>
    <name>yarn.nodemanager.local-dirs</name>
    <value>/scratch/mahidhar/13124736/mapred_scratch</value>
    <description>The hostname of the RM.
        Default: ${hadoop.tmp.dir}/nm-local-dir</description>
  </property>

<!-- yarn.nodemanager.log-dirs defaults to ${yarn.log.dir}/userlogs, where
     yarn.log.dir is set by yarn-env.sh via the YARN_LOG_DIR environment
     variable -->

<!-- these are necessary for mapreduce to work with YARN -->
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
    <description>The valid service name should only contain a-zA-Z0-9_ and can
        not start with numbers.  Default: none</description>
  </property>

<!--
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce_shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
    <description>Java class to handle the shuffle stage of
        mapreduce.
        Default:  org.apache.hadoop.mapred.ShuffleHandler</description>
  </property>
-->


</configuration>
```

yarn-site.xml parameters

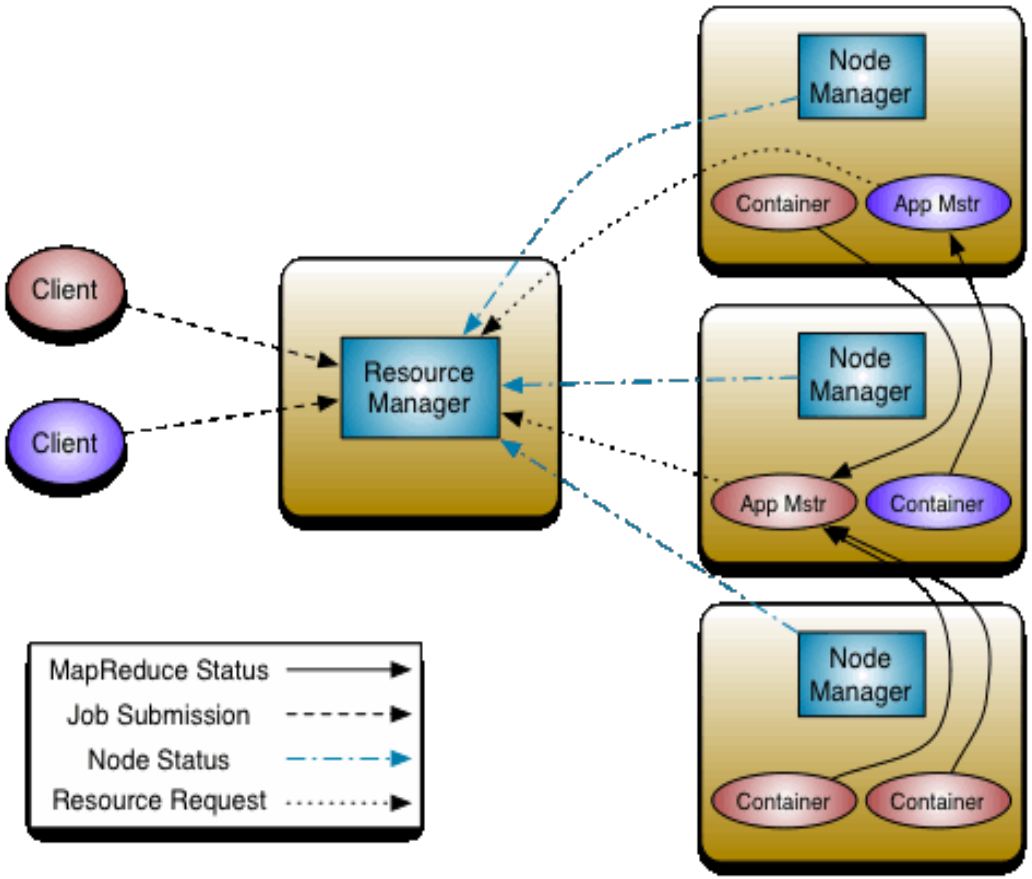
Parameter	Default	Description/Options
yarn.resourcemanager.client.thread-count	50	Number of threads used to handle applications manager requests
yarn.scheduler.minimum-allocation-mb	1024	The minimum allocation for every container request at the RM, in MBs
yarn.resourcemanager.ha.enabled	false	Enable RM high-availability

Lot more options. See:
https://hadoop.apache.org/docs/r2.7.2/hadoop-yarn/hadoop-yarn-common/yarn-default.xml

YARN: NexGen MapReduce

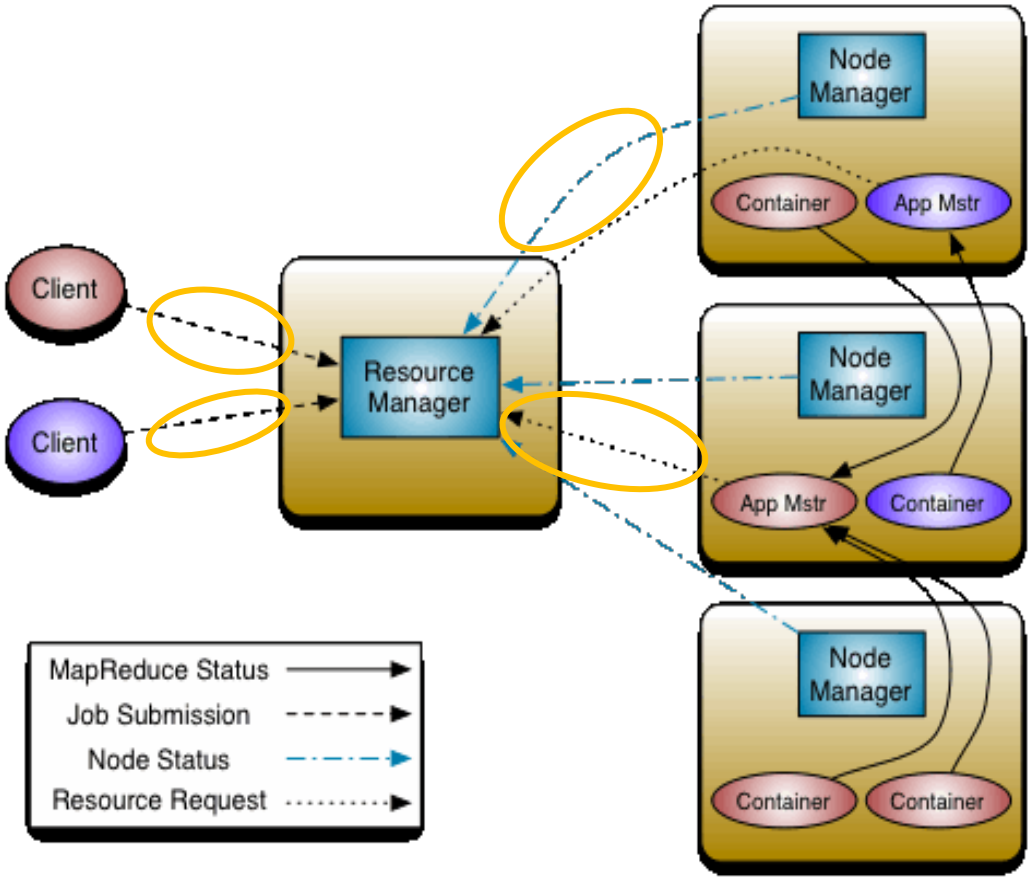
- ***Main idea – Separate resource management and job scheduling/monitoring.***
- ***Global ResourceManager (RM)***
- ***NodeManager on each node***
- ***ApplicationMaster – one for each application***

YARN Architecture



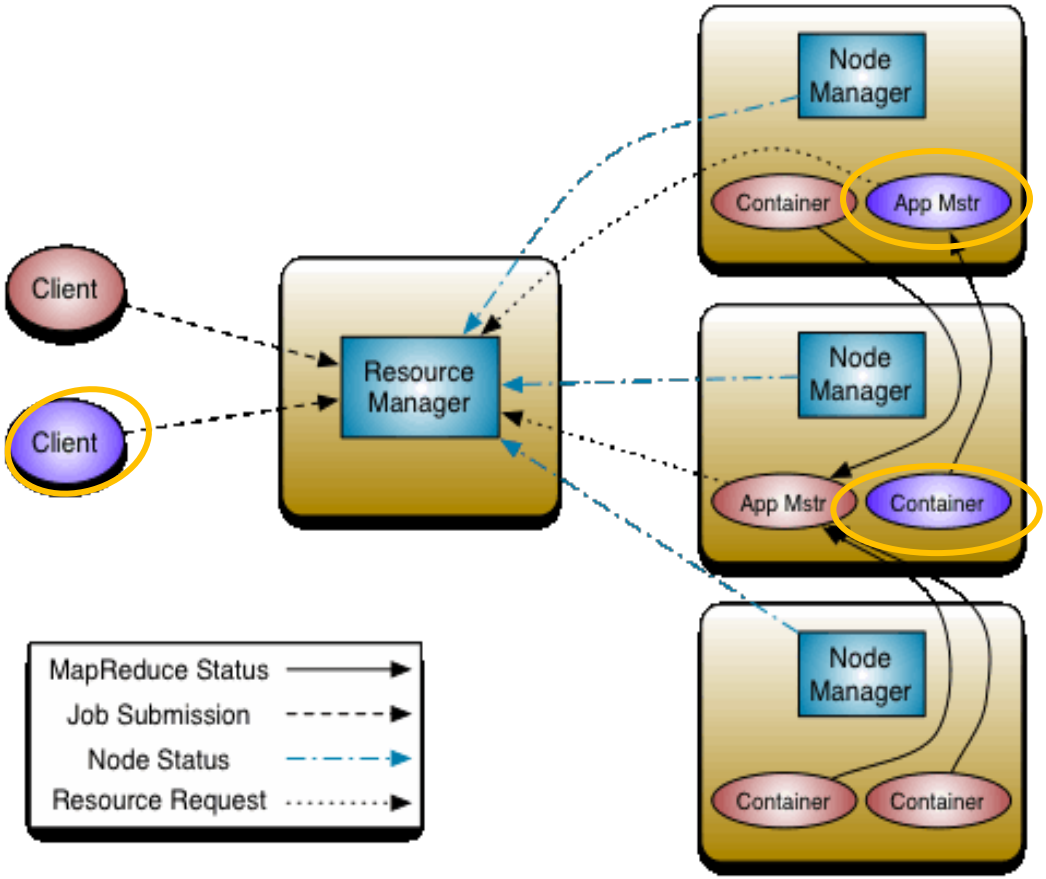
http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

YARN Architecture



Ref http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

YARN Architecture



Additional YARN Features

- *High Availability ResourceManager*
- *Timeline Server*
- *Use of Cgroups*
- *Secure Containers*
- *YARN – web services REST APIs*

Original HDFS Design

- **Single NameNode -** a master server that manages the file system namespace and regulates access to files by clients.
- **Multiple DataNodes –** typically one per node in the cluster. Functions:
  - **Manage storage - blocks of data**
  - **Serving read/write requests from clients**
  - **Block creation, deletion, replication based on instructions from NameNode**

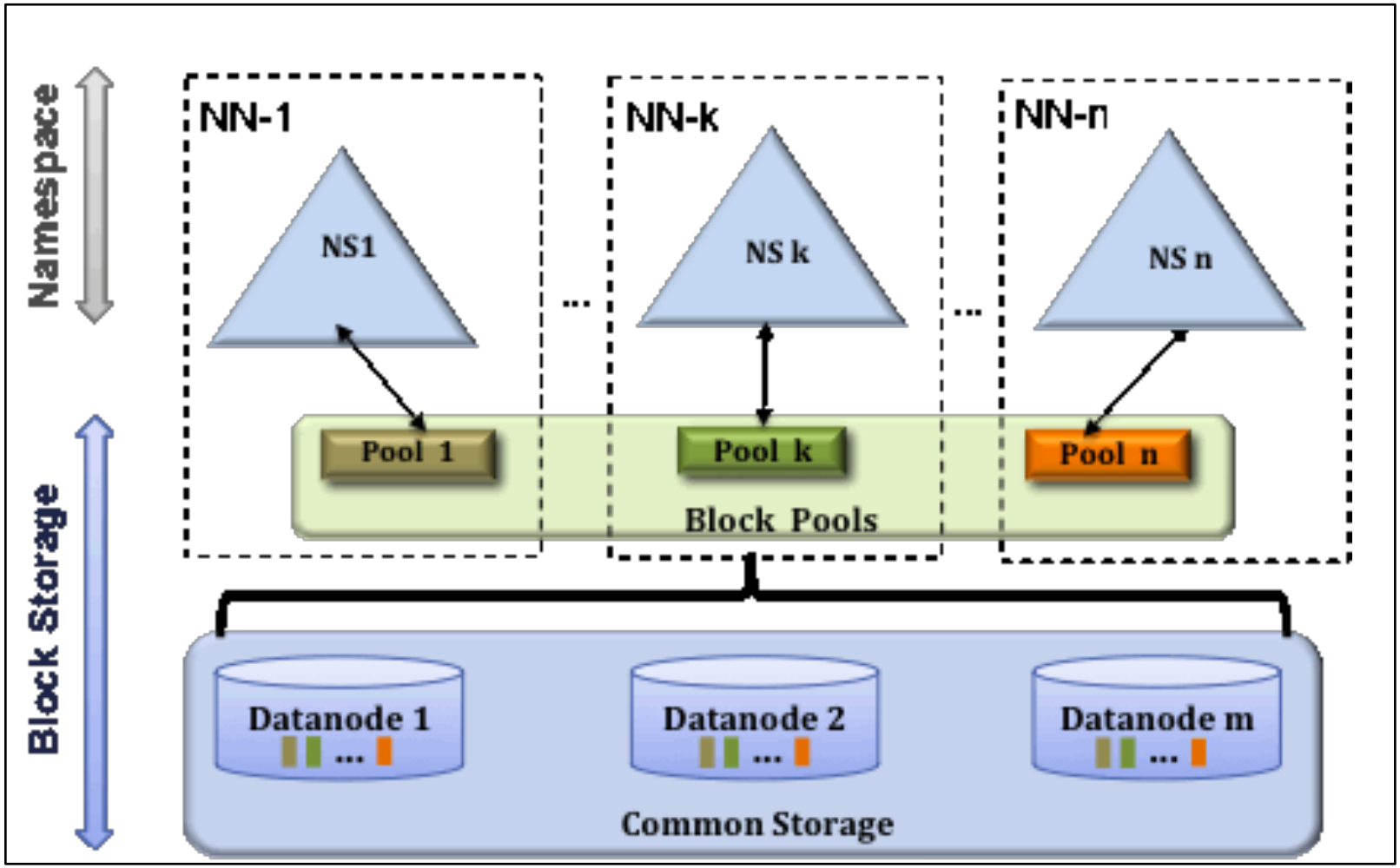
HDFS in Hadoop 2

- ***HDFS Federation***
  - *Benefits:*
    - *Increased namespace scalability*
    - *Performance*
    - *Isolation*
- ***How its done:***
  - *Multiple Namenode servers*
  - *Multiple namespaces*
  - *Block pools: each namespace has pool of blocks. DataNode will support pools from different namespaces.*

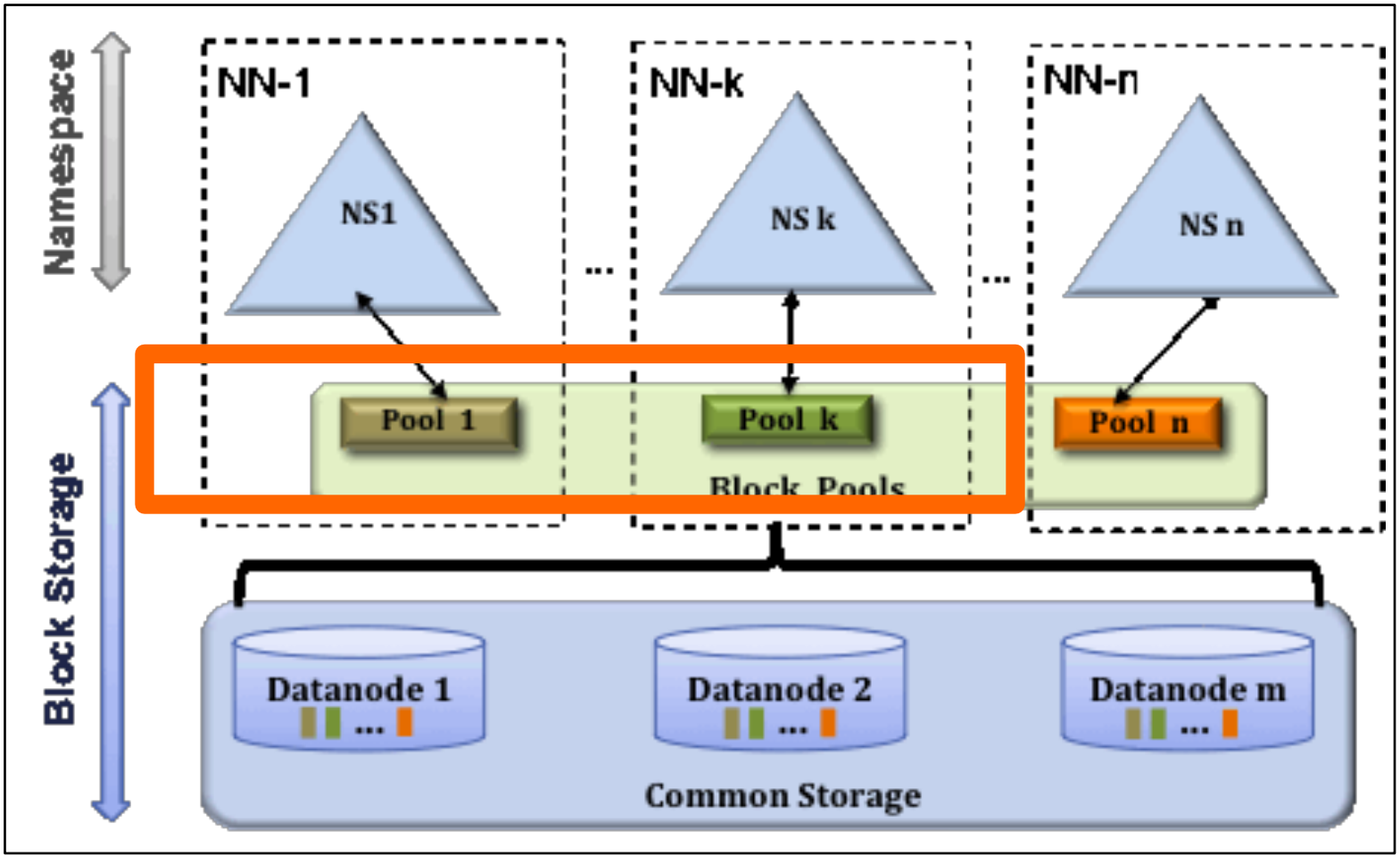
HDFS in Hadoop 2

- ***High Availability – redundant NameNodes***
- ***Heterogeneous Storage and Archival Storage***
  - *ARCHIVE, DISK, SSD, RAM_DISK*

Federation


Namespace
Block Storage
NN-1
NN-k
NN-n
NS1
NS k
NS n
Pool 1
Pool k
Pool n
Block Pools
Datanode 1
Datanode 2
Datanode m
Common Storage

Federation: Block Pools



hdfs dfsadmin

```
Usage: hdfs dfsadmin
Note: Administrative commands can only be run as the HDFS superuser.
        [-report [-live] [-dead] [-decommissioning]]
        [-safemode <enter | leave | get | wait>]
        [-saveNamespace]
        [-rollEdits]
        [-restoreFailedStorage true|false|check]
        [-refreshNodes]
        [-setQuota <quota> <dirname>...<dirname>]
        [-clrQuota <dirname>...<dirname>]
        [-setSpaceQuota <quota> <dirname>...<dirname>]
        [-clrSpaceQuota <dirname>...<dirname>]
        [-finalizeUpgrade]
        [-rollingUpgrade [<query|prepare|finalize>]]
        [-refreshServiceAcl]
        [-refreshUserToGroupsMappings]
        [-refreshSuperUserGroupsConfiguration]
        [-refreshCallQueue]
        [-refresh <host:ipc_port> <key> [arg1..argn]
        [-reconfig <datanode|...> <host:ipc_port> <start|status>]
        [-printTopology]
        [-refreshNamenodes datanode_host:ipc_port]
        [-deleteBlockPool datanode_host:ipc_port blockpoolId [force]]
        [-setBalancerBandwidth <bandwidth in bytes per second>]
        [-fetchImage <local directory>]
        [-allowSnapshot <snapshotDir>]
        [-disallowSnapshot <snapshotDir>]
        [-shutdownDatanode <datanode_host:ipc_port> [upgrade]]
        [-getDatanodeInfo <datanode_host:ipc_port>]
        [-metasave filename]
        [-setStoragePolicy path policyName]
        [-getStoragePolicy path]
        [-help [cmd]]
```

hdfs dfsadmin

```
[mahidhar@comet-10-02 apache-mahout-distribution-0.13.0]$ hdfs dfsadmin -report
17/12/05 19:57:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
Configured Capacity: 898674065408 (836.96 GB)
Present Capacity: 898558592000 (836.85 GB)
DFS Remaining: 897553924096 (835.91 GB)
DFS Used: 1004667904 (958.13 MB)
DFS Used%: 0.11%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-------------------------------------------------
Live datanodes (4):

Name: 10.22.253.172:50010 (comet-10-19.ibnet)
Hostname: comet-10-19.sdsc.edu
Decommission Status : Normal
Configured Capacity: 224670777344 (209.24 GB)
DFS Used: 237511680 (226.51 MB)
Non DFS Used: 14015488 (13.37 MB)
DFS Remaining: 224419250176 (209.01 GB)
DFS Used%: 0.11%
DFS Remaining%: 99.89%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Dec 05 19:57:54 PST 2017
```

```
Name: 10.22.253.156:50010 (comet-10-35.ibnet)
Hostname: comet-10-35.sdsc.edu
Decommission Status : Normal
Configured Capacity: 224668024832 (209.24 GB)
DFS Used: 217720832 (207.63 MB)
Non DFS Used: 15063040 (14.37 MB)
DFS Remaining: 224435240960 (209.02 GB)
DFS Used%: 0.10%
DFS Remaining%: 99.90%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Dec 05 19:57:53 PST 2017


Name: 10.22.253.167:50010 (comet-10-24.ibnet)
Hostname: comet-10-24.sdsc.edu
Decommission Status : Normal
Configured Capacity: 224664879104 (209.24 GB)
DFS Used: 269131776 (256.66 MB)
Non DFS Used: 8085504 (7.71 MB)
DFS Remaining: 224387661824 (208.98 GB)
DFS Used%: 0.12%
DFS Remaining%: 99.88%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Dec 05 19:57:53 PST 2017
```

hdfs fsck

```
[mahidhar@comet-10-02 apache-mahout-distribution-0.13.0]$ hdfs fsck /tmp/mahout-work-mahidhar/20news-all/alt.atheism/54258
17/12/05 20:01:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
Connecting to namenode via http://comet-10-02.ibnet:50070
FSCK started by mahidhar (auth:SIMPLE) from /10.22.253.189 for path /tmp/mahout-work-mahidhar/20news-all/alt.atheism/54258 at Tue Dec
 05 20:01:49 PST 2017
.Status: HEALTHY
 Total size:    11676 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 11676 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    3
 Average block replication:     3.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          4
 Number of racks:               1
FSCK ended at Tue Dec 05 20:01:49 PST 2017 in 4 milliseconds


The filesystem under path '/tmp/mahout-work-mahidhar/20news-all/alt.atheism/54258' is HEALTHY
```

hdfs version

```
[mahidhar@comet-10-02 apache-mahout-distribution-0.13.0]$ hdfs version
Hadoop 2.6.0
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r e3496499ecb8d220fba99dc5ed4c99c8f9e33bb1
Compiled by jenkins on 2014-11-13T21:10Z
Compiled with protoc 2.5.0
From source with checksum 18e43357c8f927c0695f1e9522859d6a
This command was run using /opt/hadoop/2.6.0/share/hadoop/common/hadoop-common-2.6.0.jar
```

Apache Mahout Algorithms

- **Collaborative Filtering with command line interfaces (CLI)**
- **Classification with CLI**
- **Clustering with CLI**
- **Dimensionality Reduction - Lanczos, SVD, PCA, QR**
  - Mostly via Mahout Math-Scala Core Library and Scala Domain Specific Languages (DSL)
- **Topic Models**

Mahout Classification

- **Naive Bayes**
  - Multinomial Naive Bayes
  - Transform Weight-normalized Naive Bayes
- **Hidden Markov Models**
  - Evaluation - Forward algorithm
  - Decoding - Viterbi algorithm
  - Learning - Baum-Welch algorithm
  - Usage example:

    mahout baumwelch -i hmm-input -o hmm-model -nh 3 -no 4 -e .0001 -m 1000
- **Logistic Regression using Stochastic Gradient Descent (SGD)**
  - Vector encoding package
  - SGD learning package
  - evolutionary optimization system

Mahout Classification (Contd.)

- **Random Forest**
  - Mapreduce implementation called Partial Decision Forests
  - Each mapper builds a subset of the forest using only the data available in its partition
  - Large datasets as long as each partition can be loaded in-memory.

Example: Twenty Newsgroups Classfication

- **Dataset:**
  - 20,000 newsgroup documents
  - partitioned across 20 different newsgroups
  - Example uses Mahout CBayes classifier to create a model that would classify a new document into one of the 20 newsgroups.

Reference:
https://mahout.apache.org/users/classification/twenty-newsgroups.html

Example: Twenty Newsgroups Classfication

- **Script available to do the following:**
  - Create a working directory for the dataset and all input/output.
  - Download and extract the *20news-bydate.tar.gz* from the dataset to the working directory.
  - Convert the full 20 newsgroups dataset into a < Text, Text > SequenceFile.
  - Convert and preprocesses the dataset into a < Text, VectorWritable > SequenceFile containing term frequencies for each document.
  - Split the preprocessed dataset into training and testing sets.
  - Train the classifier.
  - Test the classifier.

Example Output

```
[mahidhar@comet-10-02 apache-mahout-distribution-0.13.0]$ ./examples/bin/classif
y-20newsgroups.sh
Discovered Hadoop v2.
Setting dfs command to /opt/hadoop/2.6.0/bin/hdfs dfs, dfs rm to /opt/hadoop/2.6
.0/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. cnaivebayes-MapReduce
2. naivebayes-MapReduce
3. cnaivebayes-Spark
4. naivebayes-Spark
5. sgd
6. clean-- cleans up the work area in /tmp/mahout-work-mahidhar
Enter your choice : 1
ok. You chose 1 and we'll use cnaivebayes-MapReduce
```

- **Choose option - cnaivebayes-MapReduce**

Example Output (Data download)

```
creating work directory at /tmp/mahout-work-mahidhar
Downloading 20news-bydate
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 13.7M  100 13.7M    0     0  11.4M      0  0:00:01  0:00:01 --:--:-- 12.9M
Extracting...
```

```
+ echo 'Copying 20newsgroups data to Hadoop 2 HDFS'
Copying 20newsgroups data to Hadoop 2 HDFS
+ /opt/hadoop/2.6.0/bin/hdfs dfs -put /tmp/mahout-work-mahidhar/20news-all /tmp/
mahout-work-mahidhar/
```

Example Output

```
Creating sequence files from 20newsgroups data
+ ./bin/mahout seqdirectory -i /tmp/mahout-work-mahidhar/20news-all -o /tmp/maho
ut-work-mahidhar/20news-seq -ow
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /opt/hadoop/2.6.0/bin/hadoop and HADOOP_CONF_DIR=/home/
mahidhar/cometcluster
MAHOUT-JOB: /oasis/scratch/comet/mahidhar/temp_project/Mahout/apache-mahout-dist
ribution-0.13.0/mahout-examples-0.13.0-job.jar
17/12/05 19:30:26 INFO AbstractJob: Command line arguments: {--charset=[UTF-8],
--chunkSize=[64], --endPhase=[2147483647], --fileFilterClass=[org.apache.mahout.
text.PrefixAdditionFilter], --input=[/tmp/mahout-work-mahidhar/20news-all], --ke
yPrefix=[], --method=[mapreduce], --output=[/tmp/mahout-work-mahidhar/20news-seq
], --overwrite=null, --startPhase=[0], --tempDir=[temp]}
```

Example Output (MapReduce Job)

```
17/12/05 19:30:29 INFO YarnClientImpl: Submitted application application_1512530
453370_0001
17/12/05 19:30:29 INFO Job: The url to track the job: http://comet-10-02.ibnet:8
088/proxy/application_1512530453370_0001/
17/12/05 19:30:29 INFO Job: Running job: job_1512530453370_0001
17/12/05 19:30:34 INFO Job: Job job_1512530453370_0001 running in uber mode : fa
lse
17/12/05 19:30:34 INFO Job:  map 0% reduce 0%
17/12/05 19:30:44 INFO Job:  map 29% reduce 0%
17/12/05 19:30:47 INFO Job:  map 44% reduce 0%
17/12/05 19:30:50 INFO Job:  map 63% reduce 0%
17/12/05 19:30:53 INFO Job:  map 88% reduce 0%
17/12/05 19:30:55 INFO Job:  map 100% reduce 0%
17/12/05 19:30:55 INFO Job: Job job_1512530453370_0001 completed successfully
17/12/05 19:30:55 INFO Job: Counters: 30
```

Example Output

```
Converting sequence files to vectors
+ ./bin/mahout seq2sparse -i /tmp/mahout-work-mahidhar/20news-seq -o /tmp/mahout
-work-mahidhar/20news-vectors -lnorm -nv -wt tfidf
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /opt/hadoop/2.6.0/bin/hadoop and HADOOP_CONF_DIR=/home/
mahidhar/cometcluster
MAHOUT-JOB: /oasis/scratch/comet/mahidhar/temp_project/Mahout/apache-mahout-dist
ribution-0.13.0/mahout-examples-0.13.0-job.jar
```

```
17/12/05 19:31:03 INFO YarnClientImpl: Submitted application application_1512530
453370_0002
17/12/05 19:31:03 INFO Job: The url to track the job: http://comet-10-02.ibnet:8
088/proxy/application_1512530453370_0002/
17/12/05 19:31:03 INFO Job: Running job: job_1512530453370_0002
17/12/05 19:31:12 INFO Job: Job job_1512530453370_0002 running in uber mode : fa
lse
17/12/05 19:31:12 INFO Job:  map 0% reduce 0%
17/12/05 19:31:19 INFO Job:  map 100% reduce 0%
17/12/05 19:31:19 INFO Job: Job job_1512530453370_0002 completed successfully
```

Output (Complimentary Result)

```
17/12/05 19:36:00 INFO TestNaiveBayesDriver: Complementary Results:
```

```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :      11239       98.8739%
Incorrectly Classified Instances        :        128        1.1261%
Total Classified Instances              :      11367

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
478     0       0       0       0       0       0       0       0       0       0       0       0       0       1       0       0       0       0       2       |  481       a     = alt.atheism
0       562     0       3       0       1       1       0       0       0       0       1       1       0       0       0       0       1       0       0       |  570       b     = comp.graphics
0       4       548     14      0       2       1       0       1       0       0       2       1       0       0       0       0       0       0       1       |  574       c     = comp.os.ms-windows.misc
1       0       0       591     2       0       3       0       0       0       0       0       0       0       1       0       0       0       0       0       |  598       d     = comp.sys.ibm.pc.hardware
0       0       0       0       588     0       0       0       0       0       0       1       1       0       0       0       0       1       0       0       |  591       e     = comp.sys.mac.hardware
0       1       2       2       0       592     1       0       0       0       0       0       0       0       1       0       0       0       0       1       |  600       f     = comp.windows.x
0       0       0       1       0       0       575     3       1       0       0       1       3       0       0       0       0       1       0       0       |  585       g     = misc.forsale
0       0       0       0       0       0       0       596     2       0       0       1       1       0       0       0       0       0       0       0       |  600       h     = rec.autos
0       0       0       0       0       0       1       0       590     0       0       0       0       0       1       0       0       0       0       0       |  592       i     = rec.motorcycles
0       0       0       0       0       0       0       0       0       611     2       0       0       0       0       0       0       0       0       0       |  613       j     = rec.sport.baseball
0       0       0       0       0       0       0       0       0       1       589     0       0       0       0       0       0       0       0       0       |  590       k     = rec.sport.hockey
0       0       0       0       0       0       0       0       0       0       0       570     0       0       0       0       0       0       1       0       |  571       l     = sci.crypt
0       0       0       4       1       0       2       1       0       0       0       0       582     0       3       1       0       0       1       0       |  595       m     = sci.electronics
0       0       0       0       0       0       0       0       0       0       0       0       1       609     0       0       0       0       0       0       |  610       n     = sci.med
0       0       0       0       0       0       0       0       0       0       0       0       0       1       603     0       0       0       0       0       |  604       o     = sci.space
1       0       0       0       0       0       0       0       1       0       0       0       0       1       0       605     0       0       0       1       |  609       p     = soc.religion.christian
0       0       1       0       0       0       0       0       0       0       0       2       0       0       0       0       567     0       0       0       |  570       q     = talk.politics.guns
1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       581     0       0       |  582       r     = talk.politics.mideast
0       0       0       0       0       0       0       0       0       0       0       1       0       0       0       0       6       0       446     0       |  453       s     = talk.politics.misc
10      0       0       0       0       0       1       0       0       0       0       0       0       0       2       4       3       2       1       356     |  379       t     = talk.religion.misc

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.9802
Accuracy                                   98.8739%
Reliability                                94.0745%
Reliability (standard deviation)            0.2161
Weighted precision                          0.9888
Weighted recall                             0.9887
Weighted F1 score                           0.9887

17/12/05 19:36:00 INFO MahoutDriver: Program took 19267 ms (Minutes: 0.32111666666666666)
```

Output (Complimentary Results)

17/12/05 19:36:27 INFO TestNaiveBayesDriver: Complementary Results:

```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :      11239       98.8739%
Incorrectly Classified Instances        :        128        1.1261%
Total Classified Instances              :      11367

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
478     0       0       0       0       0       0       0       0       0       0       0       0       0       1       0       0       0       0       2       |  481        a     = alt.atheism
0       562     0       3       0       1       1       0       0       0       0       1       1       0       0       0       0       1       0       0       |  570        b     = comp.graphics
0       4       548     14      0       2       1       0       1       0       0       2       1       0       0       0       0       0       0       1       |  574        c     = comp.os.ms-windows.misc
1       0       0       591     2       0       3       0       0       0       0       0       0       0       1       0       0       0       0       0       |  598        d     = comp.sys.ibm.pc.hardware
0       0       0       0       588     0       0       0       0       0       0       1       1       0       0       0       0       1       0       0       |  591        e     = comp.sys.mac.hardware
0       1       2       2       0       592     1       0       0       0       0       0       0       0       1       0       0       0       0       1       |  600        f     = comp.windows.x
0       0       0       1       0       0       575     3       1       0       0       1       3       0       0       0       0       1       0       0       |  585        g     = misc.forsale
0       0       0       0       0       0       0       596     2       0       0       1       1       0       0       0       0       0       0       0       |  600        h     = rec.autos
0       0       0       0       0       0       1       0       590     0       0       0       0       0       1       0       0       0       0       0       |  592        i     = rec.motorcycles
0       0       0       0       0       0       0       0       0       611     2       0       0       0       0       0       0       0       0       0       |  613        j     = rec.sport.baseball
0       0       0       0       0       0       0       0       0       1       589     0       0       0       0       0       0       0       0       0       |  590        k     = rec.sport.hockey
0       0       0       0       0       0       0       0       0       0       0       570     0       0       0       0       0       1       0       0       |  571        l     = sci.crypt
0       0       0       4       1       0       2       1       0       0       0       0       582     0       3       1       0       0       1       0       |  595        m     = sci.electronics
0       0       0       0       0       0       0       0       0       0       0       0       1       609     0       0       0       0       0       0       |  610        n     = sci.med
0       0       0       0       0       0       0       0       0       0       0       0       0       1       603     0       0       0       0       0       |  604        o     = sci.space
1       0       0       0       0       0       0       0       1       0       0       0       0       1       0       605     0       0       0       1       |  609        p     = soc.religion.christian
0       0       1       0       0       0       0       0       0       0       0       2       0       0       0       0       567     0       0       0       |  570        q     = talk.politics.guns
1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       581     0       0       |  582        r     = talk.politics.mideast
0       0       0       0       0       0       0       0       0       0       0       0       1       0       0       0       0       0       6       0       446     0       |  453        s     = talk.politics.misc
10      0       0       0       0       0       1       0       0       0       0       0       0       0       2       4       3       2       1       356     |  379        t     = talk.religion.misc

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.9802
Accuracy                                   98.8739%
Reliability                                94.0745%
Reliability (standard deviation)            0.2161
Weighted precision                          0.9888
Weighted recall                             0.9887
Weighted F1 score                           0.9887

17/12/05 19:36:00 INFO MahoutDriver: Program took 19267 ms (Minutes: 0.32111666666666666)
```

Mahout Clustering Algorithms

- **K-Means clustering**
  - data directory contains multiple input files of SequenceFile
  - clusters directory contains one or more SequenceFiles(Text, Cluster) containing *k* initial clusters or canopies
  - KMeans clusters and Canopy canopies may be used for the initial clusters.
- **Canopy clustering**
  - Input Hadoop SequenceFiles containing multidimensional points
  - Canopy generation and optionally clustering

K-Means Usage

```
bin/mahout kmeans \
    -i <input vectors directory> \
    -c <input clusters directory> \
    -o <output working directory> \
    -k <optional number of initial clusters to sample from input vectors> \
    -dm <DistanceMeasure> \
    -x <maximum number of iterations> \
    -cd <optional convergence delta. Default is 0.5> \
    -ow <overwrite output directory if present>
    -cl <run input vector clustering after computing Canopies>
    -xm <execution method: sequential or mapreduce>
```

Canopy Usage

```
bin/mahout canopy \
    -i <input vectors directory> \
    -o <output working directory> \
    -dm <DistanceMeasure> \
    -t1 <T1 threshold> \
    -t2 <T2 threshold> \
    -t3 <optional reducer T1 threshold> \
    -t4 <optional reducer T2 threshold> \
    -cf <optional cluster filter size (default: 0)> \
    -ow <overwrite output directory if present>
    -cl <run input vector clustering after computing Canopies>
    -xm <execution method: sequential or mapreduce>
```

Mahout Recommender Example

- Example run on SDSC's Comet cluster
- GroupLens Movie Dataset (ml-100k.zip)
- SIMILARITY COOCURRENCE

Reference: https://chimpler.wordpress.com/2013/02/20/playing-with-the-mahout-recommendation-engine-on-a-hadoop-cluster/

SIMILARITY COOCURRENCE

- **Two movies are "similar" if they often appear together in users' rating.**
- Recommender finds 10 movies most similar to the movies the user has rated.
- Mahout computes the recommendations by running several Hadoop mapreduce jobs.
- Output in the slides is from a run on SDSC's Comet machine with a dynamically spun up Hadoop Cluster.

ml-100k.zip data

- **u.data**: contains several tuples(user_id, movie_id, rating, timestamp)
- **u.user**: contains several tuples(user_id, age, gender, occupation, zip_code)
- **u.item**: contains several tuples(movie_id, title, release_date, video_release_data, imdb_url, cat_unknown, cat_action, cat_adventure, cat_animation, cat_children, cat_comedy, cat_crime, cat_documentary, cat_drama, cat_fantasy, cat_film_noir, cat_horror, cat_musical, cat_mystery, cat_romance, cat_sci_fi, cat_thriller, cat_war, cat_western)

Steps for recommendation example

- hadoop fs -put u.data u.data
- hadoop jar $MAHOUT_JAR org.apache.mahout.cf.taste.hadoop.item.RecommenderJob -s SIMILARITY_COOCCURRENCE --input u.data --output output
- hadoop fs -getmerge output output.txt
- Post process with python script

Sample output

```
Mahidhars-MacBook-Pro:recommend mahidhar$ more output.txt
1       [527:5.0,199:5.0,660:5.0,529:5.0,200:5.0,134:5.0,66:5.0,132:5.0,265:5.0,133:5.0]
2       [200:5.0,300:5.0,197:5.0,527:5.0,134:5.0,196:5.0,132:5.0,98:5.0,265:5.0,135:5.0]
3       [137:5.0,284:5.0,248:4.8705034,14:4.8125,13:4.7894735,508:4.721683,124:4.7045455,319:4.674033,311:4.646154,272:4.6460032]
4       [100:5.0,265:5.0,1022:5.0,263:5.0,98:5.0,197:5.0,132:5.0,264:5.0,330:5.0,990:5.0]
5       [527:5.0,230:5.0,98:5.0,265:5.0,231:5.0,229:5.0,99:5.0,132:5.0,197:5.0,200:5.0]
6       [527:5.0,200:5.0,428:5.0,265:5.0,526:5.0,230:5.0,99:5.0,659:5.0,660:5.0,196:5.0]
7       [465:5.0,70:5.0,134:5.0,201:5.0,462:5.0,268:5.0,66:5.0,197:5.0,129:5.0,260:5.0]
8       [265:5.0,231:5.0,100:5.0,197:5.0,527:5.0,230:5.0,99:5.0,132:5.0,98:5.0,200:5.0]
9       [265:5.0,231:5.0,100:5.0,197:5.0,527:5.0,230:5.0,132:5.0,660:5.0,98:5.0,200:5.0]
10      [98:5.0,527:5.0,659:5.0,660:5.0,265:5.0,230:5.0,66:5.0,528:5.0,428:5.0,526:5.0]
11      [98:5.0,527:5.0,264:5.0,100:5.0,197:5.0,265:5.0,66:5.0,132:5.0,660:5.0,692:5.0]
12      [527:5.0,692:5.0,100:5.0,265:5.0,230:5.0,229:5.0,99:5.0,660:5.0,197:5.0,200:5.0]
13      [659:5.0,197:5.0,264:5.0,330:5.0,660:5.0,67:5.0,198:5.0,132:5.0,528:5.0,265:5.0]
14      [98:5.0,527:5.0,659:5.0,660:5.0,197:5.0,265:5.0,66:5.0,132:5.0,428:5.0,692:5.0]
15      [230:5.0,196:5.0,98:5.0,265:5.0,134:5.0,200:5.0,132:5.0,100:5.0,197:5.0,299:5.0]
16      [229:5.0,134:5.0,660:5.0,527:5.0,231:5.0,692:5.0,66:5.0,132:5.0,265:5.0,196:5.0]
```

- **Not human understandable!**
- **Lets correlate to the movie names**
- **Python script to do this is on reference site.**

Sample Output

```
Mahidhars-MacBook-Pro:recommend mahidhar$ python recommend.py 17 u.data u.item output.txt
Reading Movies Descriptions
Reading Rated Movies
Reading Recommendations
Rated Movies
------------------------
Mighty Aphrodite (1995), rating=3
People vs. Larry Flynt, The (1996), rating=3
Willy Wonka and the Chocolate Factory (1971), rating=4
Dead Man Walking (1995), rating=3
Toy Story (1995), rating=4
Jerry Maguire (1996), rating=2
Devil's Own, The (1997), rating=2
Phenomenon (1996), rating=1
Michael Collins (1996), rating=3
Rock, The (1996), rating=3
Swingers (1996), rating=5
Big Night (1996), rating=4
English Patient, The (1996), rating=3
Liar Liar (1997), rating=4
Dante's Peak (1997), rating=1
Courage Under Fire (1996), rating=2
City of Lost Children, The (1995), rating=4
Spitfire Grill, The (1996), rating=4
Twelve Monkeys (1995), rating=4
Fargo (1996), rating=4
Truth About Cats & Dogs, The (1996), rating=3
Trainspotting (1996), rating=4
Jungle2Jungle (1997), rating=1
Sleepers (1996), rating=1
Breaking the Waves (1996), rating=2
Full Monty, The (1997), rating=4
Leaving Las Vegas (1995), rating=4
Star Trek: First Contact (1996), rating=3
------------------------
Recommended Movies
------------------------
Bound (1996), score=4.5742574
Reservoir Dogs (1992), score=4.3958335
Donnie Brasco (1997), score=4.3333335
Clerks (1994), score=4.327869
Kingpin (1996), score=4.325843
Taxi Driver (1976), score=4.3186812
Primal Fear (1996), score=4.2922373
Beavis and Butt-head Do America (1996), score=4.2606635
This Is Spinal Tap (1984), score=4.206897
Rumble in the Bronx (1995), score=4.1976047
------------------------
```

```
------------------------
Recommended Movies
------------------------
Bound (1996), score=4.5742574
Reservoir Dogs (1992), score=4.3958335
Donnie Brasco (1997), score=4.3333335
Clerks (1994), score=4.327869
Kingpin (1996), score=4.325843
Taxi Driver (1976), score=4.3186812
Primal Fear (1996), score=4.2922373
Beavis and Butt-head Do America (1996), score=4.2606635
This Is Spinal Tap (1984), score=4.206897
Rumble in the Bronx (1995), score=4.1976047
------------------------
```

Summary

- Hadoop configuration controlled by xml configuration files. Some important ones - *core-site.xml, hdfs-site.xml, yarn-site.xml*
- Configuration parameters include HDFS replication, block sizes, security options, YARN scheduler options etc.
- Apache Mahout has a rich set of algorithms - can backend to Hadoop MapReduce, Spark, Flink etc.
- With Hadoop most Mahout runs lead to multiple Map Reduce jobs.