Large Scale Data Processing
CSE3025

Dr. Ramesh Ragala

School of Computer Science and Engineering
VIT Chennai

February 22, 2021

1 Hadoop Commands

- **Working with HDFS**
  - **To view the initial directory content**
    - **hdfs dfs -ls**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 15:05 /kmeans.ja
va
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

  - **To create a directory**
    - **hdfs dfs -mkdir directoryName**
    - Example: hdfs dfs -mkdir /ramesh
    - check the result in UI also

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -mkdir /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
Found 2 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 15:05 /kmeans.ja
va
drwxr-xr-x   - ragalayathvisra supergroup          0 2021-02-22 15:20 /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

- **Working with HDFS**
  - **Inserting data into HDFS**
    - **put** command → Copy single source, or multiple sources from local file system to the destination file system (HDFS).
    - It also reads input from stdin and writes to destination file system if the source is set to "-".
    - Copying fails if the file already exists, unless the -f flag is given.
    - **hdfs dfs -put SoureLocationFullPath DestionationLocationFullPath**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -put /home/ragalayathvisra/Downloads
/kmeans.java /ramesh/kmeans.java
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 15:31 /ramesh/kmeans.java
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -put /home/ragalayathvisra/Downloads
/kmeans.java /ramesh/kmeans.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 2 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 15:31 /ramesh/kmeans.java
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 15:33 /ramesh/kmeans.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -put - /ramesh/kmeans.txt
VIT Chennai School of Computer Science and Engineering
Ragala Ramesh, Associate Professor
Vandalur Kelambakkam Road
^Cragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup        116 2021-02-22 22:00 /ramesh/kmeans.txt._COPY
ING_
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -put -d - /ramesh/kmeans10.txt
VIT Chennai
Ramesh Ragala
Associate Professor
SCOPE
^Cragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

  - check the result in UI also

- **Working with HDFS**
  - **Inserting data into HDFS**
    - **copyFromLocal** command → Similar to the -**put** command, except that the source is restricted to a local file reference.
    - **hdfs dfs -copyFromLocal LocalSoureLocationFullPath DestionationLocationFullPath**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -copyFromLocal /home/ragalayathvisra
/Downloads/kmeans.java /ramesh/kmeans120.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:15 /ramesh/kmeans120.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

  - check the result in UI also

- **Working with HDFS**
  - **Removing the files from HDFS**
    - **rm** command → Delete files which are passes as arguments.
    - If trash is enabled, file system instead moves the deleted file to a trash directory
    - The trash feature is disabled by default.
    - **hdfs dfs -rm DestionationLocationFullPathInHDFS**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:21 /ramesh/kmeans120.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -rm /ramesh/kmeans120.txt
Deleted /ramesh/kmeans120.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 2 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:23 /ramesh/kmeans120.txt
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:23 /ramesh/kmeans121.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -rm /ramesh/*
Deleted /ramesh/kmeans120.txt
Deleted /ramesh/kmeans121.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 2 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:24 /ramesh/kmeans120.txt
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:24 /ramesh/kmeans121.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -rm /ramesh/kmeans120.txt /ramesh/km
eans121.txt
Deleted /ramesh/kmeans120.txt
Deleted /ramesh/kmeans121.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

  - check the result in UI also

- **Working with HDFS**
  - **Removing a Directory from HDFS**
    - **rmdir** command → Deletes a directory.
    - **hdfs dfs -rmdir DestionationLocationOfDirectoryFullPathInHDFS**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - ragalayathvisra supergroup          0 2021-02-22 22:26 /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -rmdir /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - ragalayathvisra supergroup          0 2021-02-22 22:40 /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /ramesh
Found 1 items
-rw-r--r--   1 ragalayathvisra supergroup       5026 2021-02-22 22:40 /ramesh/kmeans120.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -rmdir --ignore-fail-on-non-empty /r
amesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - ragalayathvisra supergroup          0 2021-02-22 22:40 /ramesh
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

  - check the result in UI also

- **Working with HDFS**
  - **Retrieving Data from HDFS**
    - **cat** command → Copies source paths to stdout.
    - **hdfs dfs -cat PathOfSourceFileInHDFS**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -cat /ramesh/kmeans120.txt
/**
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements.  See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership.  The ASF licenses this file
 * to you 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.
 */
package kmeansdemo;
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -cat hdfs://localhost:9000/ramesh/km
eans120.txt
/**
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements.  See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership.  The ASF licenses this file
 * to you 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.
 */
package kmeansdemo;
```

- **Working with HDFS**
  - **Retrieving Data from HDFS**
    - **get** command → Copy files to the local file system.
    - It is the inverse operation of put
    - **hdfs dfs -get FullPathOfSourceFileInHdfs FullPathOfDestionLocationinLocalMachine**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -get /ramesh/kmeans120.txt /home/rag
alayathvisra/Music/kmeans130.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ cd /home/ragalayathvisra/Music/
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ ls
k-means++  kmeans130.txt  LSDA_LabAttendance  LSDA_TheoryAttendance
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -get hdfs://localhost:9000/ramesh/km
eans120.txt /home/ragalayathvisra/Music/kmeans131.tzt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ cd /home/ragalayathvisra/Music/
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ ls
k-means++  kmeans130.txt  kmeans131.tzt  LSDA_LabAttendance  LSDA_TheoryAttendance
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$
```

  - check the result in UI also

- **Working with HDFS**
  - **Retrieving Data from HDFS**
    - **copyToLocal** command → Similar to get command, except that the destination is restricted to a local file reference.
    - **hdfs dfs -copyToLocal FullPathOfSourceFileInHdfs FullPathOfDestionLocationinLocalMachine**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -copyToLocal /ramesh/kmeans120.txt /
home/ragalayathvisra/Music/kmeans150.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ cd /home/ragalayathvisra/Music/
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ ls
k-means++       kmeans131.tzt  LSDA_LabAttendance
kmeans130.txt   kmeans150.txt  LSDA_TheoryAttendance
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ hdfs dfs -copyToLocal hdfs://localhost:
9000/ramesh/kmeans120.txt /home/ragalayathvisra/Music/kmeans160.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ ls
k-means++       kmeans131.tzt  kmeans160.txt       LSDA_TheoryAttendance
kmeans130.txt   kmeans150.txt  LSDA_LabAttendance
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$
```

  - check the result in UI also

- **Working with HDFS**
  - **Retrieving Data from HDFS**
    - **Count** command → Count the number of directories, files and bytes under the paths that match the specified file pattern.
    - It gets the quota and the usage.
    - The output columns with -count are: DIR_COUNT, FILE_COUNT, CONTENT_SIZE and PATHNAME.
    - **hdfs dfs -count FullPathOfSourceFileInHdfs**

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$ hdfs dfs -count /
           2            1               5026 /
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~$
```

```
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ hdfs dfs -copyToLocal hdfs://localhost:
9000/ramesh/kmeans120.txt /home/ragalayathvisra/Music/kmeans160.txt
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$ ls
k-means++       kmeans131.tzt  kmeans160.txt       LSDA_TheoryAttendance
kmeans130.txt   kmeans150.txt  LSDA_LabAttendance
ragalayathvisra@ragalayathvisra-ThinkPad-E470:~/Music$
```

  - check the result in UI also