# TOBB ETU HADOOP - IBM BigInsights Cluster Erişim ve Kullanımı

#### İrfan Bahadır KATİPOĞLU\*

#### $3 \ {\rm Mart} \ 2014$

193.140.108.162
10
10 GB
$50~\mathrm{GB}$
4
4
1.1.1
QuickStart Edition v2.1.0.1
bkz.EkA

#### 1 SSH Erişimi

MasterNode IP adresinden size verilen kullanıcı adı ve parola ile SSH erişimi kurabilirsiniz. Linux kullanıcıları SSH ile aşağıdaki gibi bağlanabilirler:

ssh etu-user1@193.140.108.162

Windows kullanıcıları ise Putty<sup>ii</sup> kullanabilirler.

### 2 SFTP Erişimi

MasterNode'a dosya yükleyip indirmek için hesaplarınıza SFTP<sup>iii</sup> ile erişebilirsiniz. FileZilla<sup>iv</sup> için örnek ekran şekilde gösterilmiştir.

<sup>`</sup>ibkatipoglu@etu.edu.tr - bahadir@bahadir.me

<sup>&</sup>lt;sup>i</sup>NameNode aynı zamanda DatNode görevi görüyor

 $<sup>^{\</sup>rm ii} \rm http://www.chiark.greenend.org.uk/\ sgtatham/putty/download.html$ 

<sup>&</sup>lt;sup>iii</sup>Secure Shell File Transfer Protocol

<sup>&</sup>lt;sup>iv</sup>https://filezilla-project.org/



(b) Putty

Şekil 1: Sunucu Bağlantısı Yapılandırma Ekranları

# 3 BigInsights Web Console Erişimi

Tarayıcıdan<sup>i</sup> http://193.140.108.162:8080/ adresini yazıp size verilen kullanıcı adı ve parola ile konsola erişebilirsiniz.

IBM© InfoSphere® BigInsights™ Quick Start Edition	
(for Non-Production Environment)	
Please enter your information	
User name:	
etu-user1	_
Password:	_
Login Cancel	
Licensed Materials - Property of IBM Corp. @ Copyright 2010, 2013. IBM copportation IBM, the IBM Logs, InfoSphers, Biplingshis, Power Systems and WebSphere are trademarks of IBM Corporation, registered in many jurisdictions worksived, auxia and all ava-based trademarks and loops are trademarks or registered trademarks of Oracle and/or its affiliates.	
armates.	

Şekil 2: Giriş ekranı

 $<sup>{\</sup>rm ^iFirefox}$ önerilir, Chrome bazı özellikleri gösteremeyebiliyor



Şekil 3: Konsol ana ekranı

/elcome Dashboard	Cluster Status Files	Applications	Application Status	BigSheets	
					Refresh Interval: 15 seconds v
Nodes	<b>Ø</b> 4				
Map/Reduce	Running	Nodes			
HDFS	Running	Add nodes	Remove nodes		
Big SQL	😋 Running		Host	Status	Roles
Catalog	Running		er applied	Anst is running	datanoda taektraekar
Hive	😵 Unavailable	slave01.etubilbi		O Host is running	datanode, tasktracker
HttpFS	😵 Running	©Running slav		O Host is running	datanode, tasktracker
Dozie Zookeeper	⊗ Running ⊘ Running	ma	ster.etubilbi	O Host is running	hive-server, secondarynamenode, zookeeper-client-port, bigsql-server, hive-web-interface, tasktracker, oozie- server, httpfs-server, jobtracker, datanode, namenode
		1.1.1.1			

Şekil 4: Cluster Ekranı

# 4 Örnek Uygulama

Örnekte Ödev1'de verilen uygulama çalıştırılacaktır. Öncelikle verilen dataset'i sunucuya yüklüyoruz. Bunun için FileZilla - SFTP aracılığı ile sisteme bağlanıyoruz. Sunucu tarafını sağ taraf göstermektedir. Bağlantı ilk sağlandığında otomatik olarak /home/<kullanıcı-adı> alanını gösteriyor olmalıdır. Şekilde kırmızı ile gösterilen alana verilen dosyayı (reutersnews.rar) bilgisayarımızdan sürükleyip bırakıyoruz. Böylece dosya kısa bir süre sonra sunucuya yüklenecektir. Hadoop üzerinde çalıştırılacak olan oluşturduğumuz jar dosyasınıda (örnekte *bigdata-hw-1.jar*) aynı şekilde sunucuya atıyoruz.



Şekil 5: Cluster Ekranı

Bundan sonraki aşamayı sunucuya SSH bağlantısı kurarak terminalden hallediyoruz.

```
# Extract dataset
$
 unrar x reutersnews.rar
 Create HDFS directories
#
$ hadoop dfs -mkdir hw1
$ hadoop dfs -mkdir hw1/input
# Put files to the HDFS.
 'time' tag is optional for measuring elapsed time.
#
$ time hadoop dfs -put reuters-news/*.txt hw1/input
        0m30.882s
real
        0m10.915s
user
        0m2.199s
sys
# Check files.
# Make sure input folder is not empty (824.537 bytes)
# and output folder is already removed.
$ hadoop dfs -du hw1
Found 1 items
824537
             2
    \u00e3 hdfs://master.etubilbi:9000/user/etu-userx/hw1/input
```

```
# Run the job
$ time hadoop jar bigdata-hw-1.jar hw1/input hw1/output
... WARN snappy.LoadSnappy: Snappy native library is 🖌
   ... INFO util.NativeCodeLoader: Loaded the native-hadoop \checkmark
   ₲ library
... INFO snappy.LoadSnappy: Snappy native library loaded
... INFO mapred.FileInputFormat: Total input paths to \checkmark
   └→ process : 1
... INFO mapred.JobClient: Running job: ∠
   └ job_201403021631_0010
... INFO mapred.JobClient: map 0% reduce 0%
. . .
Job Counters
... INFO mapred.JobClient:
                                 Data-local map tasks=2
... INFO mapred.JobClient:
                                 SLOTS_MILLIS_MAPS=11752
... INFO mapred.JobClient:
                                 Launched map tasks=2
                             Total time spent by all 2
... INFO mapred.JobClient:
    \backsim reduces waiting after reserving slots (ms)=0
... INFO mapred.JobClient: Total time spent by all 🖉
   \backsim maps waiting after reserving slots (ms)=0
... INFO mapred.JobClient: Launched reduce tasks=1
... INFO mapred.JobClient: SLOTS_MILLIS_REDUCES=11358
. . .
# Check the results.
$ hadoop dfs -ls hw1/output
Found 3 items
-rw-r--r-- ... /user/etu-userx/hw1/output/_SUCCESS
drwxr-xr-x ... /user/etu-userx/hw1/output/_logs
-rw-r--r- ... /user/etu-userx/hw1/output/part-00000
# Print out the output
$ hadoop dfs -cat hw1/output/*0
1
        2763
2
        20016
. . .
23
        2
24
        1
# We must delete the output folder recursively for another \checkmark
   5 run
$ hadoop dfs -rmr hw1/output
                     Kod 1: Ödev1 Uygulaması
```

Çalışmakta olan ve geçmiş Job'lara ait durumu Console'dan Application Status bölümünden izleyebilirsiniz.

ekome	Dashboard Ouster Status	Files Applications Ap	plication Status	BgSheets					
pplication S	Status								
cheduled W	forkflows   Workflows   Jobs							Refres	h Interval: 15 seconds 👻 🚺
Status	Name	Job	D	Map % Complete	Reducer % Complete	Start Time	<ul> <li>End</li> </ul>	Time User N	ame Priority
No filte	er applied								
1.	CountByLength	job_20140303	21631_0012	83%	27%	2014-03-03 1	4:16 N	A etu-us	er1 NORMAL
	CountByLength	job_20140303	21631_0011	100%	103%	2014-03-03 1	4:05 2014-03-	03 14:06 etu-us	er1 NORMAL
	CountByLength	job_20140303	21631_0010	100%	100%	2014-03-03 1	4:04 2014-03-	03 14:05 etu-us	er1 NORMAL
	CountByLength	job_20140303	21631_0008	100%	100%	2014-03-03 1	0:23 2014-03-	03 10:24 etu-us	er1 NORMAL
	CountByLength	job_2014030	21631_0007	100%	100%	2014-03-03 1	0:13 2014-03-	03 10:19 etu-us	er1 NORMAL
	CountByLength	job_20140303	21631_0006	100%	100%	2014-03-02 2	1:39 2014-03-	02 21:39 biadn	nin NORMAL
	CountByLength	job 2014030	21631 0006	100%	103%	2014-03-02 2	0:25 2014-03-	02 20 25 etu-us	er1 NORMAL
	CountByLength	job 2014030	21631 0004	100%	103%	2014-03-02 1	8:03 2014-03-	02 18:03 etu-us	er1 NORMAL
	CountRvl enoth	iph 2014030	21631_0002	100%	101%	2014-03-02 1	8:00 2014-03-	02 18 01 efturits	er1 NORMAI
	CountByLength	job_20140303	21631 0001	100%	100%	2014-03-02 1	7:53 2014-03-	02 17:53 biadn	nin NORMAL
		· -							
10 of 10 ite	ems		10	0   25   50   100   Al					ii ( <b>1</b> ) →
ob Counter	s Job Configuration								
Tasks									
Type	Total Tasks	Successful Tasks	Failed Ta	sks Kill	ed Tasks R	unning Tasks	Pending Tasks	Start Time	End Time
setup	2	1	0		1	0	0	NA	ONaN-NaN-NaN NaN:NaN
map	925	689	0		0	13	223	NA	N/A
reduce	1	0	0		0	1	0	ONaN-NaN-NaN NaN:NaN	N/A

Şekil 6: Uygulama Durum Ekranı

# Ekler

### A BigInsights Bileşenleri

- dm
- zookeeper
- data-compression
- scheduler
- adaptivemapred
- sftp
- text-analytics
- hadoop
- derby
- jaql
- $\bullet \ \, hive^i$
- pig
- lucene
- flume
- ei

- machine-learning
- hcatalog
- sqoop
- bigsql
- bigindex
- oozie
- orchestrator
- jaqlserver
- $\bullet \ \ {\rm console}$
- eclipsetooling
- sheets
- import-export
- httpfs
- monitoring

<sup>&</sup>lt;sup>i</sup>Geçici olarak hizmet dışı