TOBB ETU
HADOOP - IBM BigInsights
Cluster Erişim ve Kullanımı

İrfan Bahadır KATİPOĞLU*

3 Mart 2014

Master Node Harici IP	193.140.108.162
Max Kullanıcı Sayısı	10
Kullanıcı Başına Master Node Alanı Kotası	10 GB
Kullanıcı Başına HDFS Dizini Kotası	50 GB
Toplam Node Sayısı	4
Data Node Sayısı[i]	4
Hadoop Sürümü	1.1.1
BigInsights Sürümü	QuickStart Edition v2.1.0.1
Yüklü Bileşenler	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[ii] kullanabilirler.

2 SFTP Erişimi

MasterNode'a dosya yükleyip indirmek için hesaplarınıza SFTP[iii] ile erişebilirsiniz. FileZilla[iv] için örnek ekran şekilde gösterilmiştir.

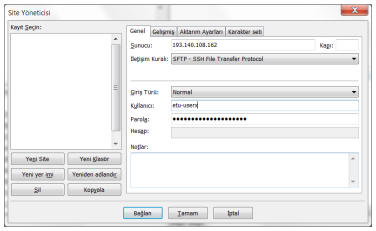
*ibkatipoglu@etu.edu.tr — bahadir@bahadir.me

[i] NameNode aynı zamanda DatNode görevi görüyor

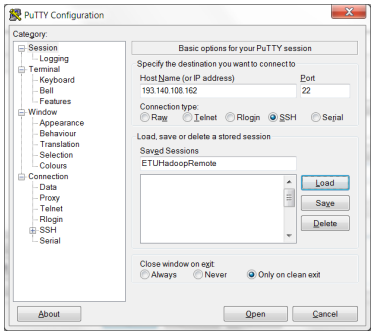
[ii] http://www.chiark.greenend.org.uk/ sgtatham/putty/download.html

[iii] Secure Shell File Transfer Protocol

[iv] https://filezilla-project.org/



(a) FileZilla

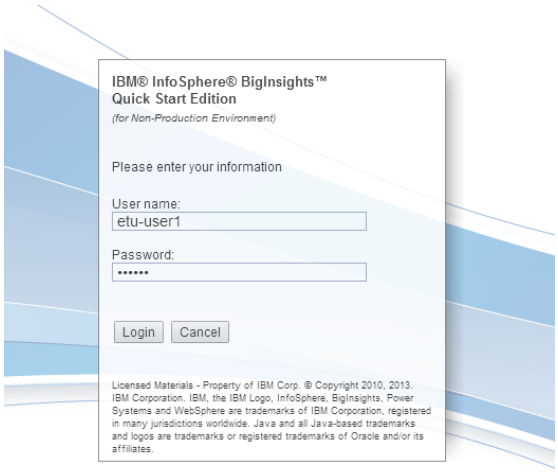


(b) Putty

Şekil 1: Sunucu Bağlantısı Yapılandırma Ekranları

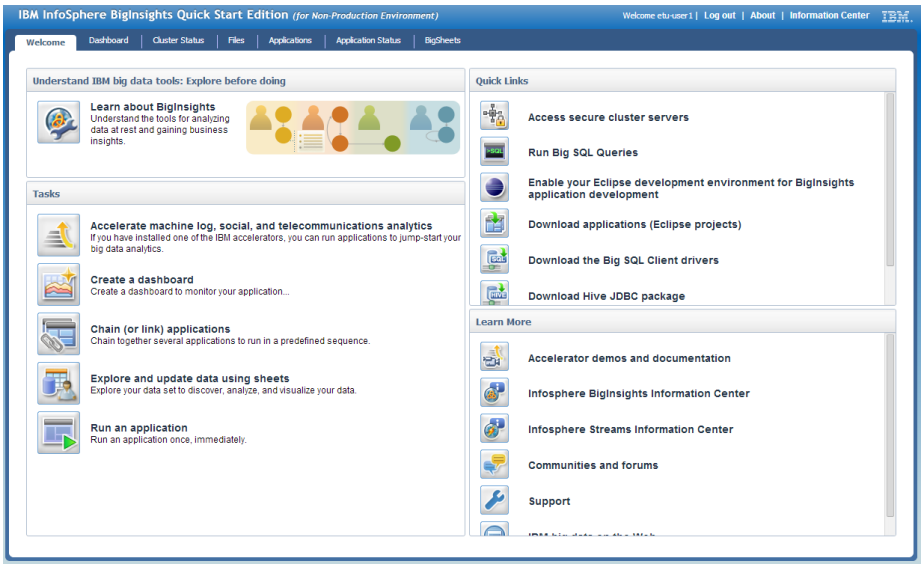
3 BigInsights Web Console Erişimi

Tarayıcıdan[i] *http://193.140.108.162:8080/* adresini yazıp size verilen kullanıcı adı ve parola ile konsola erişebilirsiniz.

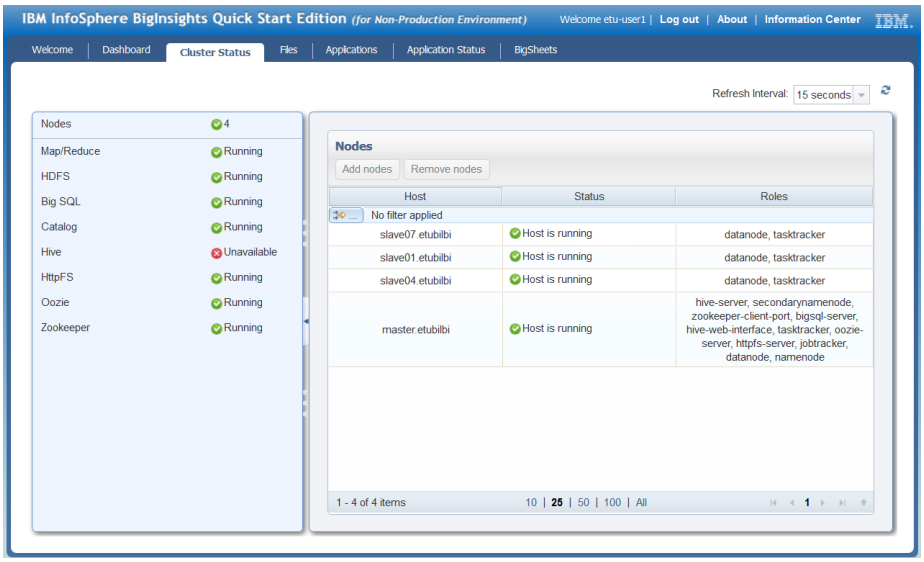


Şekil 2: Giriş ekranı

[i]Firefox önerilir, Chrome bazı özellikleri gösteremeyebiliyor



Şekil 3: Konsol ana ekranı

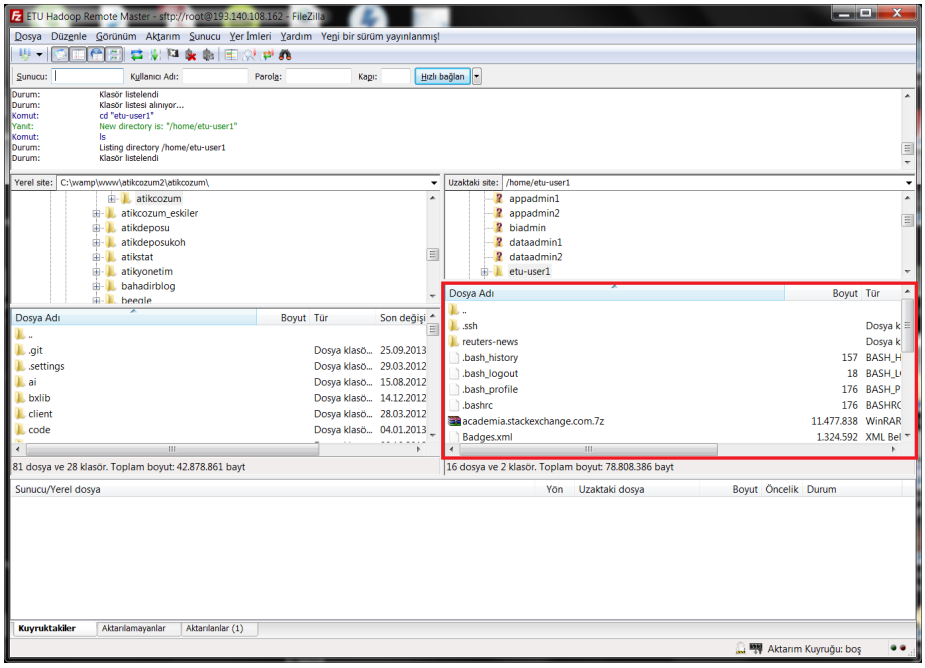


Şekil 4: Cluster Ekranı

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 dosyasımıda (örnekte *bigdata-hw-1.jar*) aynı şekilde sunucuya atıyoruz.



Şekil 5: Cluster Ekranı

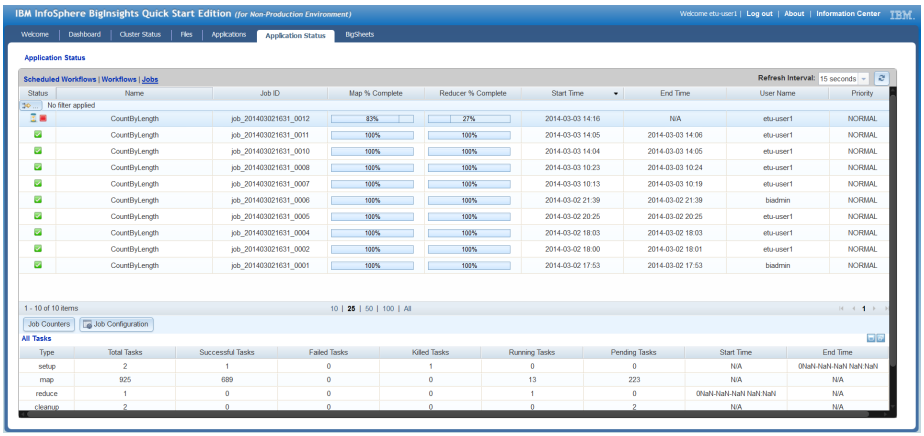
Bundan sonraki aşamayı sunucuya SSH bağlantısı kurarak terminalden halledi-yoruz.

```
# 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

real    0m30.882s
user    0m10.915s
sys     0m2.199s

# 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      ↙
    ↳ 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 ↙
    ↳ available
... INFO util.NativeCodeLoader: Loaded the native-hadoop ↙
    ↳ library
... INFO snappy.LoadSnappy: Snappy native library loaded
... INFO mapred.FileInputFormat: Total input paths to ↙
    ↳ 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
... INFO mapred.JobClient:     Total time spent by all ↙
    ↳ reduces waiting after reserving slots (ms)=0
... INFO mapred.JobClient:     Total time spent by all ↙
    ↳ 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 ↙
    ↳ run
$ hadoop dfs -rmr hw1/output
```

Kod 1: Ödev1 Uygulaması

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



Şekil 6: Uygulama Durum Ekranı

Ekler

A BigInsights Bileşenleri

- dm
- zookeeper
- data-compression
- scheduler
- adaptivemapred
- sftp
- text-analytics
- hadoop
- derby
- jaql
- hive[i]
- pig
- lucene
- flume
- ei
- machine-learning
- hcatalog
- sqoop
- bigsql
- bigindex
- oozie
- orchestrator
- jaqlserver
- console
- eclipsetooling
- sheets
- import-export
- httpfs
- monitoring

[i] Geçici olarak hizmet dışı