

# HADOOP

YASİN FİŞNE && SEFA DALGIÇ

# Konular

1.BÜYÜK VERİ  
KAVRAMI

2.HADOOP NEDİR

3.HADOOP'UN KISA  
TARİHÇESİ

4.HADOOP  
MİMARISI

5.KULLANIM ALANLARI

6.HADOOP  
KURULUMU

7.WORD-COUNT  
UYGULAMASI

NEDİR?

# **BIG DATA**

Sosyal medya paylaşımları, fotoğraf arşivleri sürekli kayıt alınan "log" dosyaları gibi farklı kaynaklardan elde edilen tüm bu verilerin anlamlı ve işlenebilir hale dönüştürülmüş biçimidir.

# Büyük Veri Bileşenleri

BÜYÜK VERİ'Yİ ANLAMAK İÇİN  
ONUN OLUŞUMUNDAKİ BEŞ  
BİLEŞENİ İNCELEMELER FAYDALI  
OLACAKTIR. BUNLAR; VARIETY,  
VELOCITY, VOLUME,  
VERIFICATION VE VALUE OLARAK  
5V ŞEKLİNDE ADLANDIRILABİLİR.

## VARIETY

Verinin geldiği kaynakların çeşitliliği (email, facebook, videolar, resimler, ses kayıtları v.s.).

---

## VELOCITY

Verinin değişim veya birikme hızı

## VOLUME

Verinin kapladığı alan

## VERIFICATION

Verideki değişimdir

## VALUE

Büyük Veri'nin veri üretim ve işleme katmanlarınızdan sonra kurum için bir artı değer yaratıyor olması gerekmektedir.

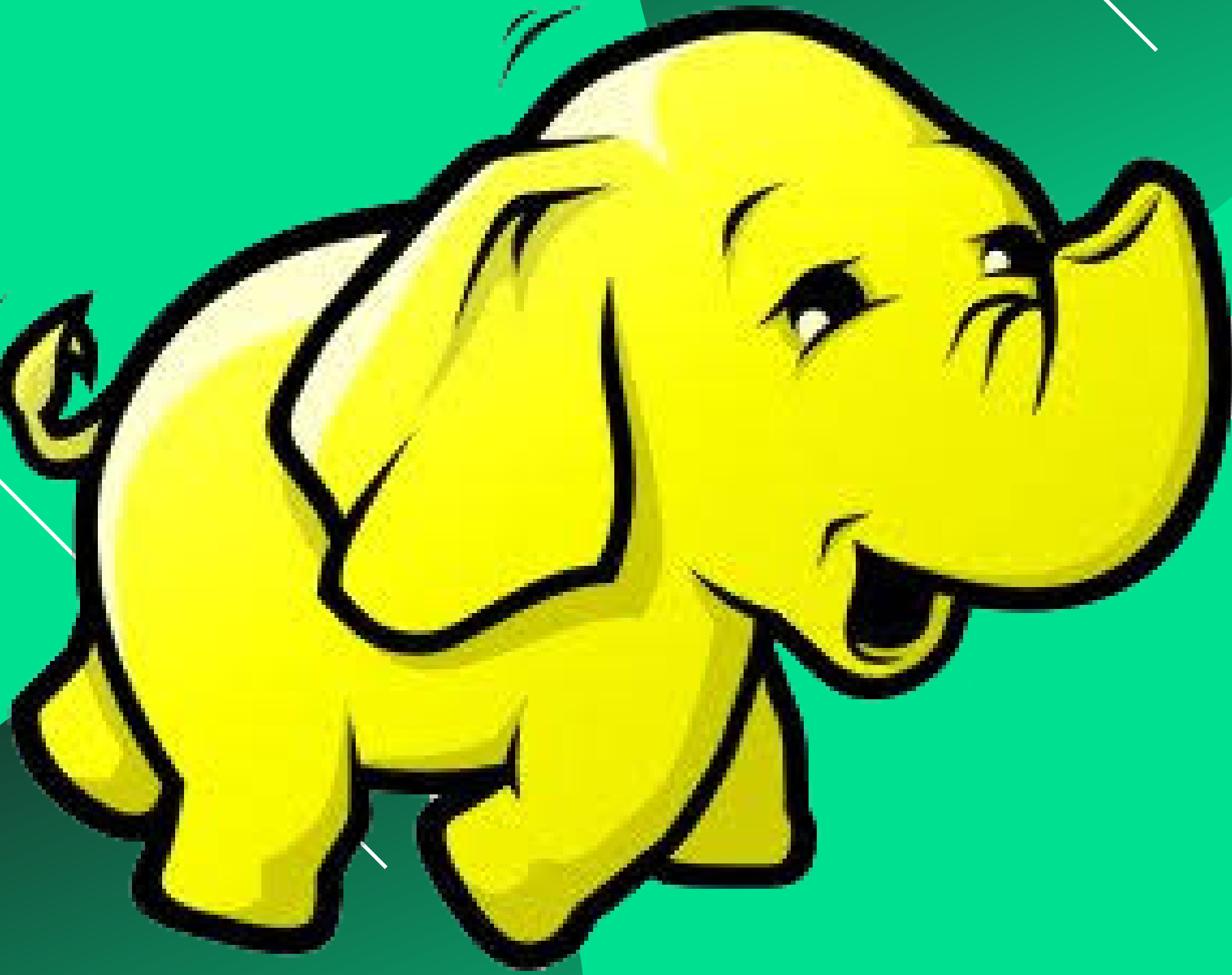
- Açık kaynak kodlu ,dağıtk,  
ölçeklenebilir, hata dayanıklı bir  
Apache projesidir.

- Map-Reduce işlemlerini hedef  
almaktadır.

- Büyük Veri dünyasında düşük  
maliyetli ve verimli çözümler  
üretir..

- Büyük ölçekteki işlemleri ve  
hesaplamaları hedefler.

HADOOP  
RENT



# Çıkış Amacı

"Kabul edilebilir zaman ve maliyet ile büyük veri üzerinde işlem yapılabilir mi?" sorusuna yanıt aramaktır.

# Tarihçesi

2005

Doug Cutting ve Michael J. Cafarella tarafından Nutch arama motoru için Hadoop geliştirildi.

Yahoo destekli projeydi

2006

Yahoo projeyi Apache Software Foundation'a transfer etti.



# Tarihçesi

2009

Avro ve Chukwa projeleri de Hadoop Framework ailesine eklendi.

2010

Hadoop'un HBase, Hive ve Pig projeleri tamamlanarak Hadoop'a daha fazla işlem gücü kazandırdılar.

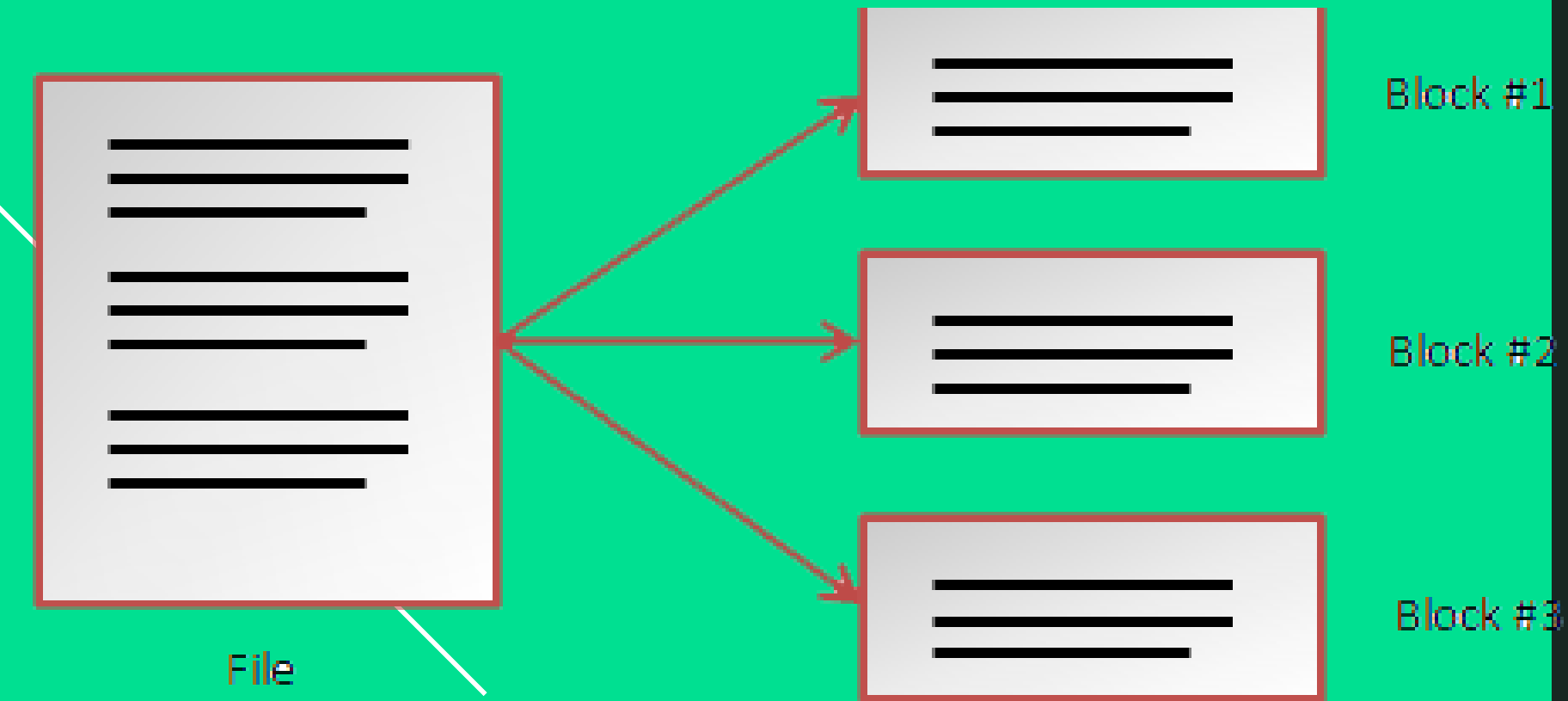
2011

ZooKeeper tamamlandı.

2013

Hadoop 1.1.2 ve Hadoop 2.0.3 alpha. – Cassandra, Mahout projeleri eklendi.

# Hadoop Büyük Verileri Nasıl Saklar ?

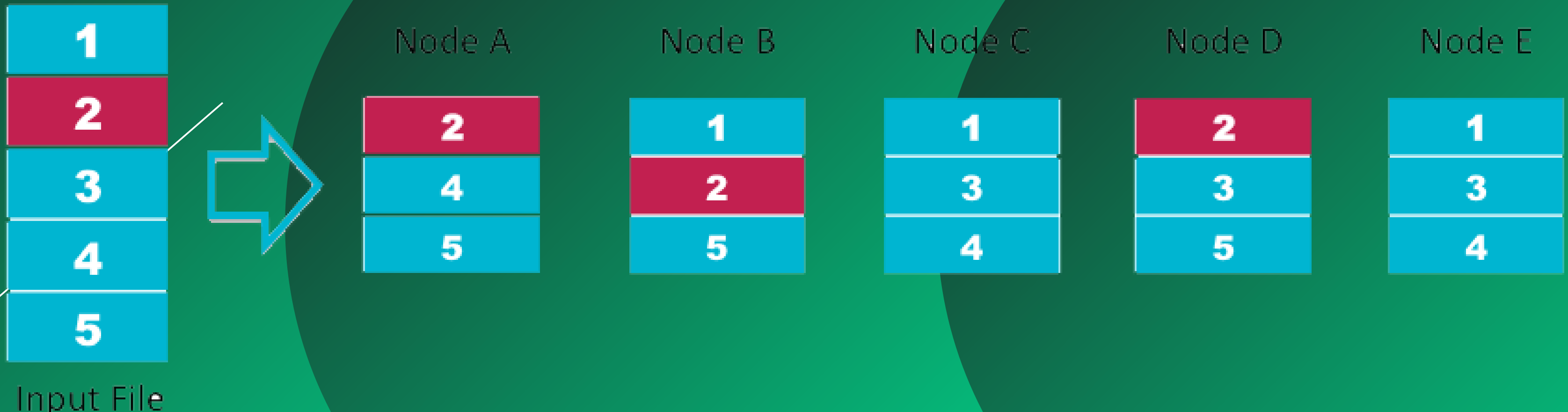


Hadoop içerisinde büyük verileri sakladığımız bileşene HDFS (Hadoop Distributed File System) denir.

Büyük verileri HDFS sistemine yüklediğimiz zaman, Hadoop bu veriler bloklara ayırır.

Farklı bloklara ayrılan veriler Hadoop Cluster üzerinde farklı node lara dağılır. Şimdilik her bir node u farklı bir makine olarak düşünebiliriz. Alttaki şekilde görüldüğü gibi Input File içerisindeki bloklar farklı node lara dağıtılmıştır . Burada dikkat etmemiz gereken en önemli hususlardan bir tanesi her bir blok çoklanarak kaydedilmiştir . Mesela 2 numaralı blok 3 farklı (Node A , Node B , Node D) node üzerine dağıtılmıştır. (Replication factor) Bunun asıl nedeni ise node lardan bir tanesi zarar gördüğünde veya sistemden çıktığında veri kaybının yaşanmasını engellemek.

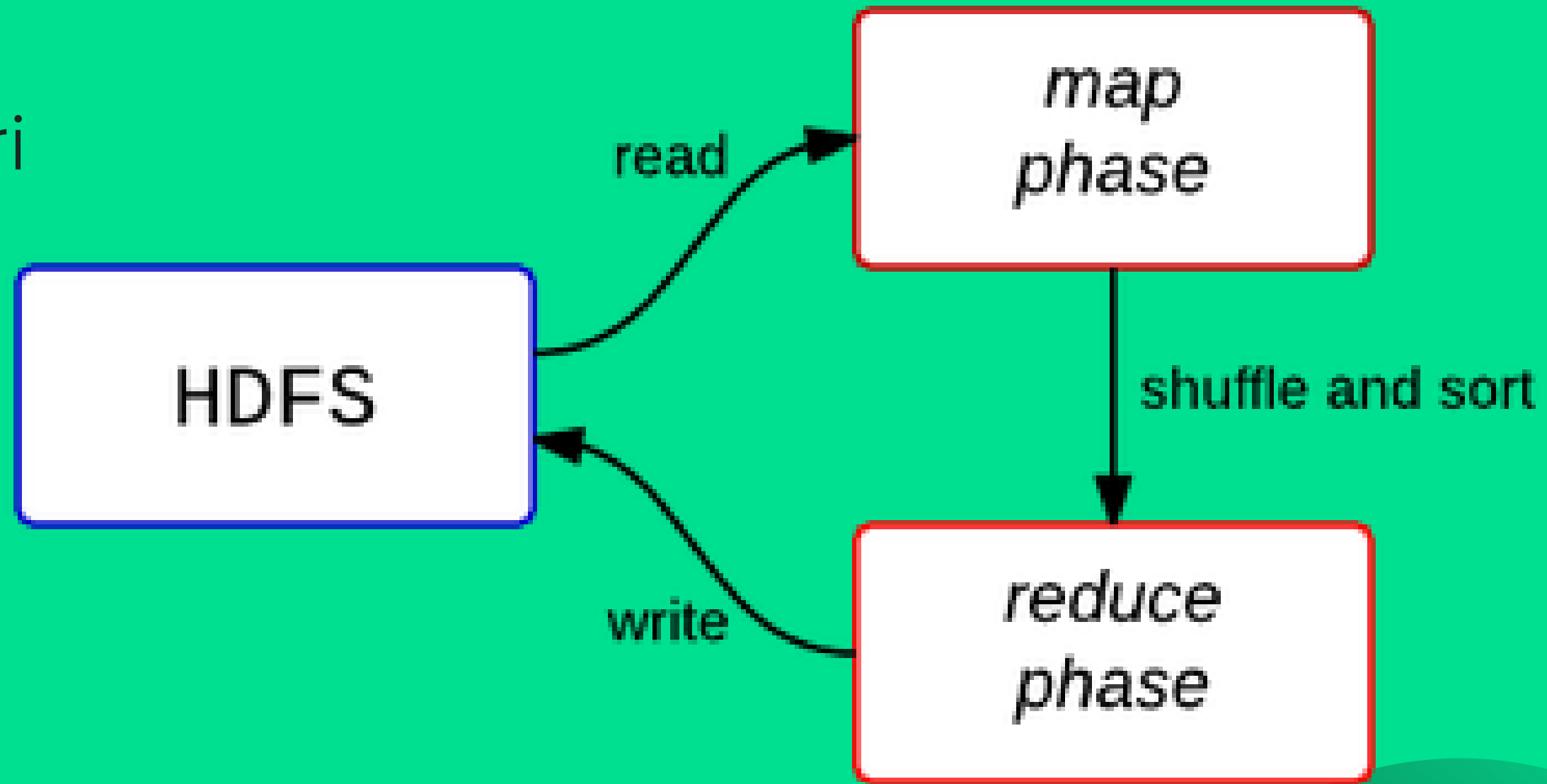
## HDFS Data Distribution



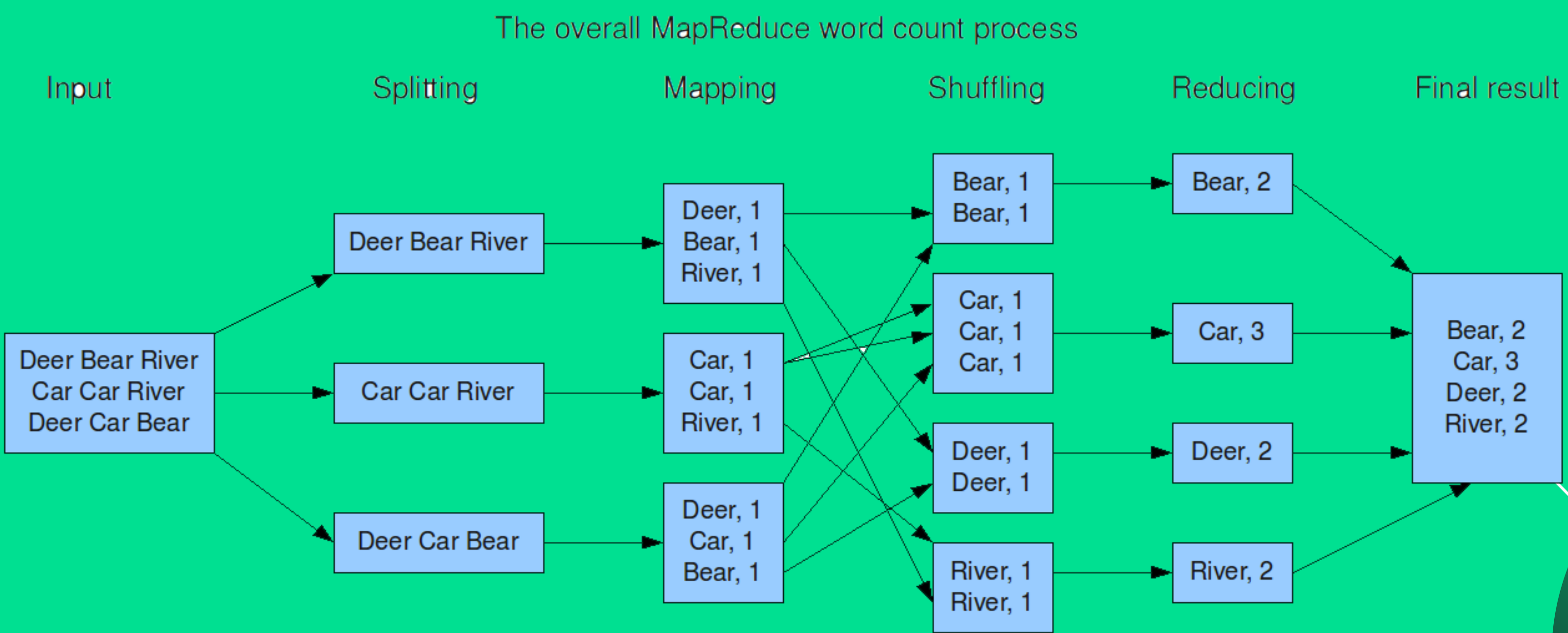
# Hadoop Verileri Paralel Olarak Nasıl İşler ?

## MAP-REDUCE NEDİR ?

Hadoop içerisinde büyük verileri paralel olarak işleyebileceğimiz bileşene MapReduce denir. Veri kümeleri HDFS üzerinden yüklendikten sonra Map ve Reduce fazları işletilir . Bu kodlamaları Java , Pig ve Hive .. ile geliştirebiliriz



Örnek olarak bir text dosyasının içerisindeki kelime sayısını bulan MapReduce programını inceleyelim . MapReduce şu adımlardan oluşacaktır ;



- ## SPLITTING

Veriler 64 MB lik bloklara ayrılır.Bu değer değiştirilebilir.

- ## MAPPING

Burada her bir kelime key(word) ve value(1) şeklinde değerlere ayrılır.

- ## SHUFFLING

Map işlemlerinden çıkan sonuç Reducer'a yönlendirilir.Amacımız Word-count uygulaması olduğu için aynı kelime grubu aynı Reducer'a yönlendirilir.

- ## REDUCING

Gelen sonuçlar üzerinden toplama işlemi yapılır ve sonuçlar istediğiniz kaynaklara yazılır.(HDFS, SQL, NoSQL)



## REKLAMLAR

Kullanıcı Davranışlarının araştırılması

## ARAMALAR

Her türlü arama, metin, kişi, otel, uçak vb.

## GÜVENLİK

Anormalliklerin anlaşılması

# Kullanım Alanları





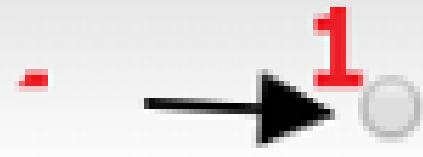
# Özet















Genel olarak özetlemek gerekirse çok yüksek trafikte akan bir veriniz olduğu zaman (Günlük 100 Milyon+) verileri HDFS üzerinde saklayabilir ve MapReduce ile verilerinizi analiz edebilirsiniz. Alternatif olarak diğer NoSQL (Mongo, ElasticSearch) saklama yöntemlerini yada ApacheSpark gibi paralel veri işleme yöntemlerini tercih edebilirsiniz.



## Java SE Development Kit 8u171

You must accept the [Oracle Binary Code License Agreement for Java SE](#) to download this software.

 ☒ Accept License Agreement ☐ Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	77.97 MB	 <a href="#">jdk-8u171-linux-arm32-vfp-hflt.tar.gz</a>
Linux ARM 64 Hard Float ABI	74.89 MB	 <a href="#">jdk-8u171-linux-arm64-vfp-hflt.tar.gz</a>
Linux x86	170.05 MB	 <a href="#">jdk-8u171-linux-i586.rpm</a>
Linux x86	184.88 MB	 <a href="#">jdk-8u171-linux-i586.tar.gz</a>
Linux x64	167.14 MB	 <a href="#">jdk-8u171-linux-x64.rpm</a>
Linux x64	182.05 MB	 <a href="#">jdk-8u171-linux-x64.tar.gz</a>
Mac OS X x64	247.84 MB	 <a href="#">jdk-8u171-macosx-x64.dmg</a>
Solaris SPARC 64-bit (SVR4 package)	139.83 MB	 <a href="#">jdk-8u171-solaris-sparcv9.tar.Z</a>
Solaris SPARC 64-bit	99.19 MB	 <a href="#">jdk-8u171-solaris-sparcv9.tar.gz</a>
Solaris x64 (SVR4 package)	140.6 MB	 <a href="#">jdk-8u171-solaris-x64.tar.Z</a>
Solaris x64	97.05 MB	 <a href="#">jdk-8u171-solaris-x64.tar.gz</a>
Windows x86	199.1 MB	 <a href="#">jdk-8u171-windows-i586.exe</a>
Windows x64	207.27 MB	 <a href="#">jdk-8u171-windows-x64.exe</a> 

Hadoop kurmadan  
önce java jdk  
kurulu olmalıdır.

<https://www.oracle.com/technetwork/java/javase/downloads/jdk13-downloads-5672538.html>



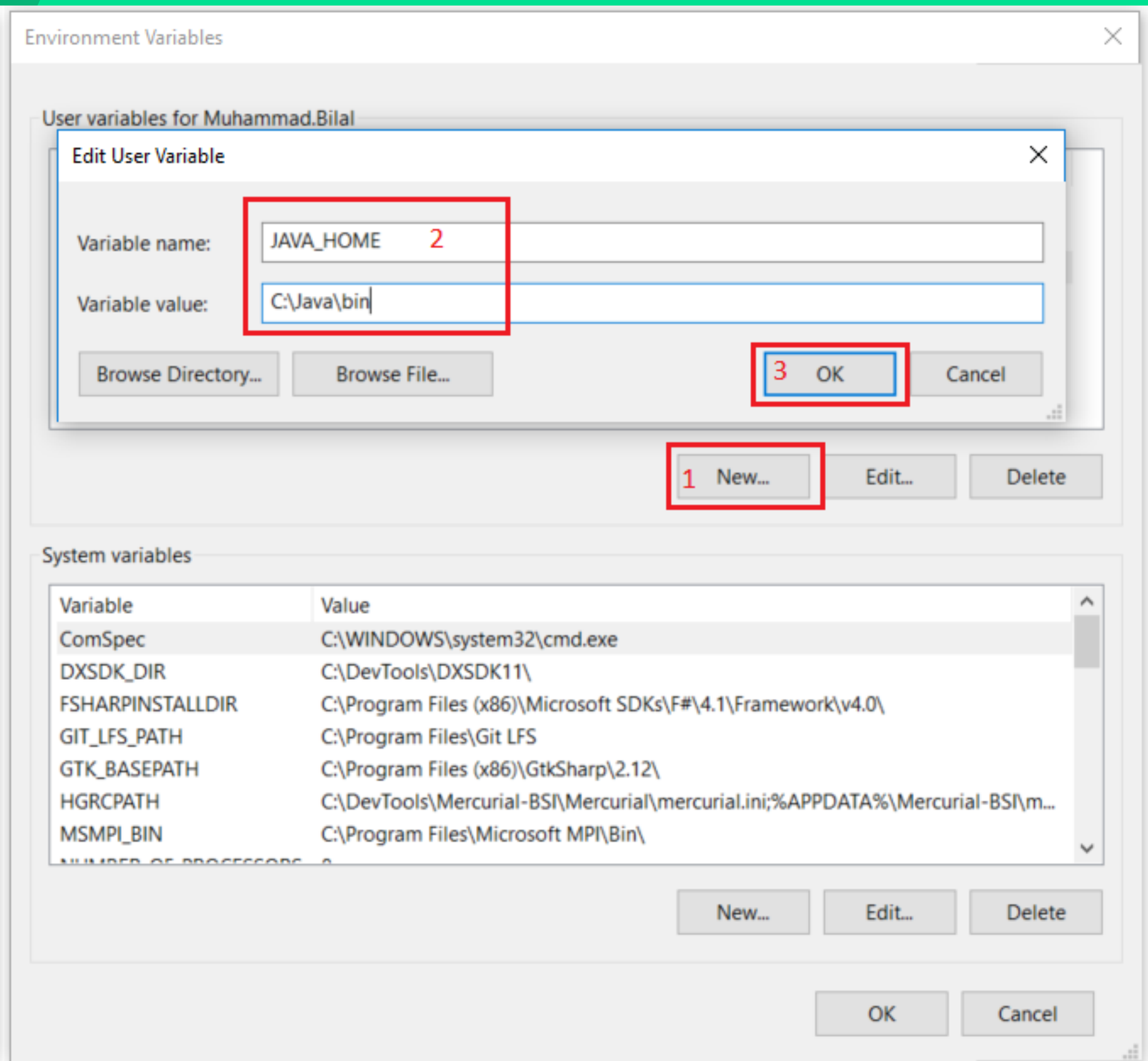
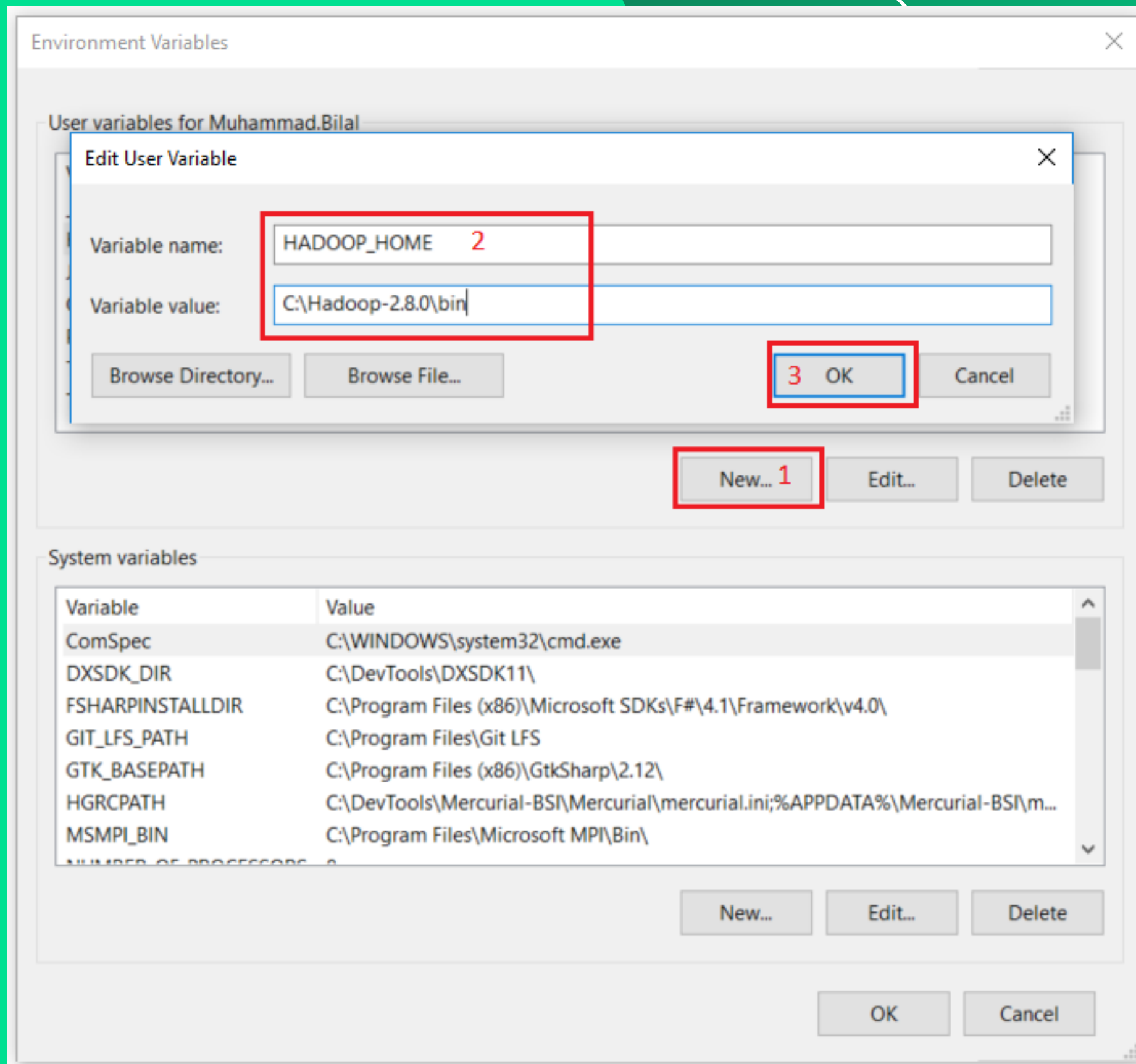
## Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Binary download	Release notes
2.10.0	2019 Oct 29	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>
3.1.3	2019 Oct 21	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>
3.2.1	2019 Sep 22	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>
3.1.2	2019 Feb 6	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>
2.9.2	2018 Nov 19	<a href="#">source (checksum signature)</a>	<a href="#">binary (checksum signature)</a>	<a href="#">Announcement</a>

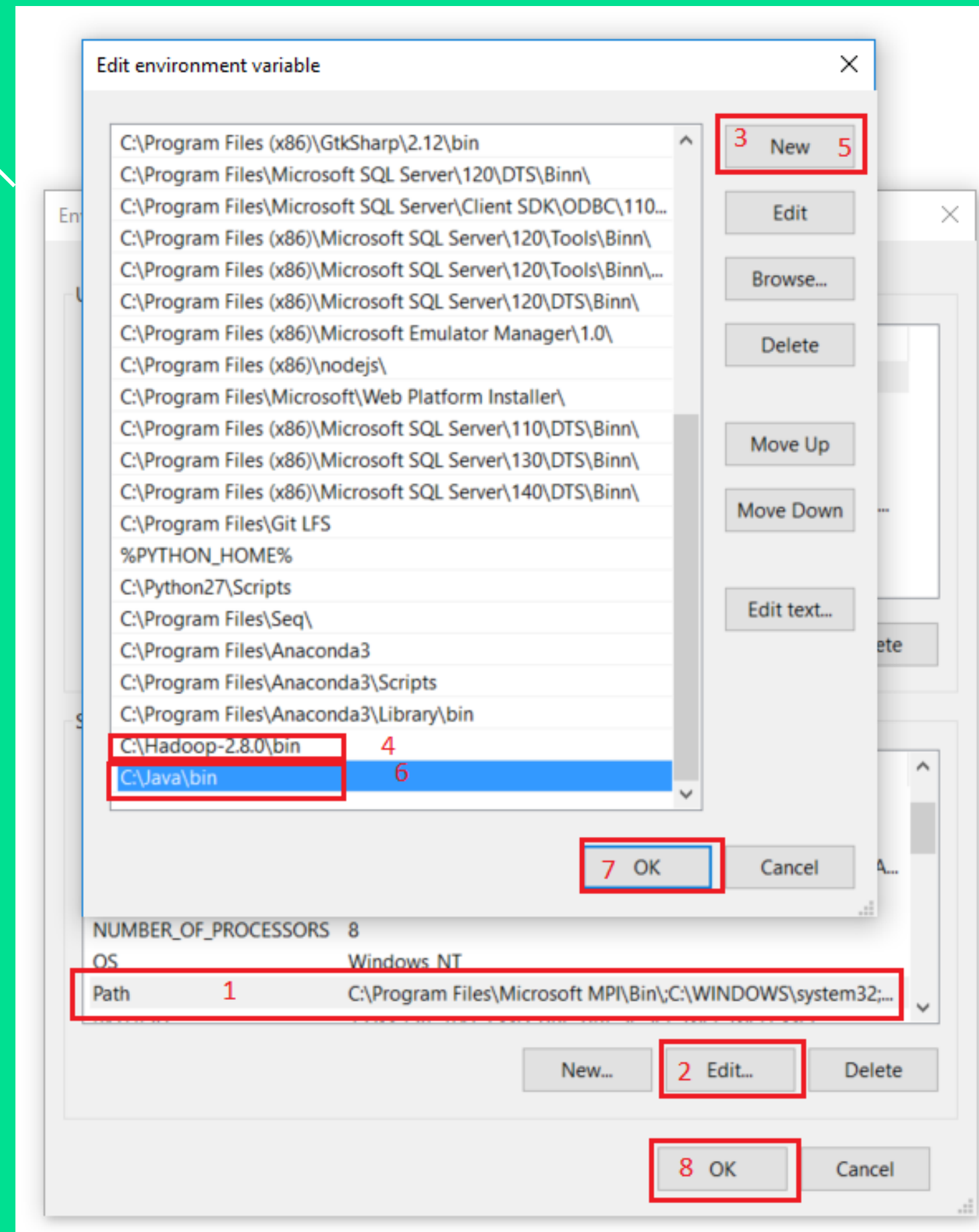
Hadoop tar.gz  
uzantılı source  
dosyası inidirilir.

<https://hadoop.apache.org/releases.html>



Sistem değişkenleri olarak Hadoop ve JDK dosyaları eklenir.

Aynı iki konumu path değeri olarak da ekliyoruz.



\hadoop\etc\hadoop  
klasöründe bulunan xml  
dosyaları üzerinde  
birtakım değişiklik  
yapmamız gerekiyor

PC > Downloads > hadoop-3.1.0 > hadoop-3.1.0 > etc > hadoop				
Name	Date modified	Type	Size	
^	3/30/2018 5:43 AM	File	2 KB	
container-executor.cfg	3/30/2018 5:31 AM	XML Document	1 KB	
core-site	3/30/2018 5:31 AM	Windows Comma...	4 KB	
hadoop-env	3/30/2018 5:31 AM	SH File	16 KB	
hadoop-env.sh	3/30/2018 5:52 AM	PROPERTIES File	4 KB	
hadoop-metrics2.properties	3/30/2018 5:31 AM	XML Document	11 KB	
hadoop-policy	3/30/2018 5:31 AM	EXAMPLE File	4 KB	
hadoop-user-functions.sh.example	3/30/2018 5:31 AM	XML Document	1 KB	
hdfs-site	3/30/2018 5:33 AM	SH File	2 KB	
httpfs-env.sh	3/30/2018 5:33 AM	PROPERTIES File	2 KB	
httpfs-log4j.properties	3/30/2018 5:33 AM	SECRET File	1 KB	
httpfs-signature.secret	3/30/2018 5:33 AM	XML Document	1 KB	
httpfs-site	3/30/2018 5:33 AM	XML Document	4 KB	
kms-acls	3/30/2018 5:31 AM	SH File	2 KB	
kms-env.sh	3/30/2018 5:31 AM	PROPERTIES File	2 KB	
kms-log4j.properties	3/30/2018 5:31 AM	XML Document	1 KB	
kms-site	3/30/2018 5:31 AM	PROPERTIES File	14 KB	
log4j.properties	3/30/2018 5:31 AM	Windows Comma...	1 KB	
mapred-env	3/30/2018 5:44 AM	SH File	2 KB	
mapred-env.sh	3/30/2018 5:44 AM	TEMPLATE File	5 KB	
mapred-queues.xml.template	3/30/2018 5:44 AM	XML Document	1 KB	
mapred-site	3/30/2018 5:44 AM	EXAMPLE File	3 KB	
ssl-client.xml.example	3/30/2018 5:31 AM	EXAMPLE File	3 KB	
ssl-server.xml.example	3/30/2018 5:31 AM	TEMPLATE File	3 KB	
user_ec_policies.xml.template	3/30/2018 5:33 AM	File	1 KB	
workers	3/30/2018 5:31 AM	Windows Comma...	3 KB	
yarn-env	3/30/2018 5:43 AM	SH File	6 KB	
yarn-env.sh	3/30/2018 5:43 AM	PROPERTIES File	3 KB	
yarnservice-log4j.properties	3/30/2018 5:43 AM	XML Document	1 KB	
yarn-site	3/30/2018 5:43 AM			



```
<configuration>  
  <property>  
    <name>fs.defaultFS</name>  
    <value>hdfs://localhost:9000</value>  
  </property>  
</configuration>
```

ilk olarak core-site.xml dosyasına bu kodları ekliyoruz.

```
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>
```

Daha sonra mapred-site.xml dosyasına bu kodları ekliyoruz.

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/hadoop-2.8.0/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/hadoop-2.8.0/data/datanode</value>
  </property>
</configuration>
```

Daha sonra Hadoop klasörüne data klasörü ve data klasörünün içine de datanode ve namenode diye 2 tane klasör ekliyoruz ve ardından bu hdfs-site.xml dosyasına bu kodları ekliyoruz.



```
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
</configuration>
```

yarn-site.xml dosyasına bu kodları ekliyoruz.

```
@rem The java implementation to use.  Required.  
@rem set JAVA_HOME=%JAVA_HOME%  
set JAVA_HOME=C:\java
```

hadoop-env.cmd dosyasına java dosyasının konumunu set etmemiz gerekiyor.

C:\ Komut İstemi

```
Microsoft Windows [Version 10.0.18362.476]  
(c) 2019 Microsoft Corporation. Tüm hakları saklıdır.
```

```
C:\Users\sefap>hdfs namenode -format
```

Cmd üzerinden "hdfs namenode -format" komutunu yazıyoruz.

Select C:\WINDOWS\system32\cmd.exe

```
C:\>cd Hadoop-2.8.0\sbin
```

```
C:\Hadoop-2.8.0\sbin>start-all.cmd
```

```
This script is Deprecated. Instead use start-dfs.cmd  
starting yarn daemons
```


```
C:\Hadoop-2.8.0\sbin>
```

```
Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - yarn nodemanager

17/07/20 15:50:09 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:12 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:15 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:18 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:21 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:24 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:27 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:30 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:33 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:36 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:39 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:42 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:46 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:49 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:52 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:55 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:50:58 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:01 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:04 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:07 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:10 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:13 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:16 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:19 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:22 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:25 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:29 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:32 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
17/07/20 15:51:35 WARN util.SysInfoWindows: Expected split length of sysInfo to be 11. Got 7
```

Ardından start-all.cmd komutunu giriyoruz ve 4 yeni komut sistemi açılıp çalışıyor.

← → ↻ ⓘ localhost:8088/cluster ☆ ⋮

 **All Applications** Logged in as: dr.who

▼ Cluster

- About
- Nodes
- Node Labels
- Applications
  - NEW
  - NEW SAVING
  - SUBMITTED
  - ACCEPTED
  - RUNNING
  - FINISHED
  - FAILED
  - KILLED
- Scheduler

► Tools

**Cluster Metrics**

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

**Cluster Nodes Metrics**

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

**Scheduler Metrics**

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 ▼ entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																	

Showing 0 to 0 of 0 entries First Previous Next Last

Ve son olarak tarayıcıya `http://localhost:8088` açıyoruz. Eğer yukarıdaki gibi bir ekran görüyorsanız kurulum başarıyla tamamlanmıştır.

# Bizi dinlediğiniz için teşekkürler

YASIN FIŞNE & SEFA DALGIÇ