SISTEM OPERASI UNTUK PEMROSESAN BIG DATA DENGAN BERBASIS CENTOS 7

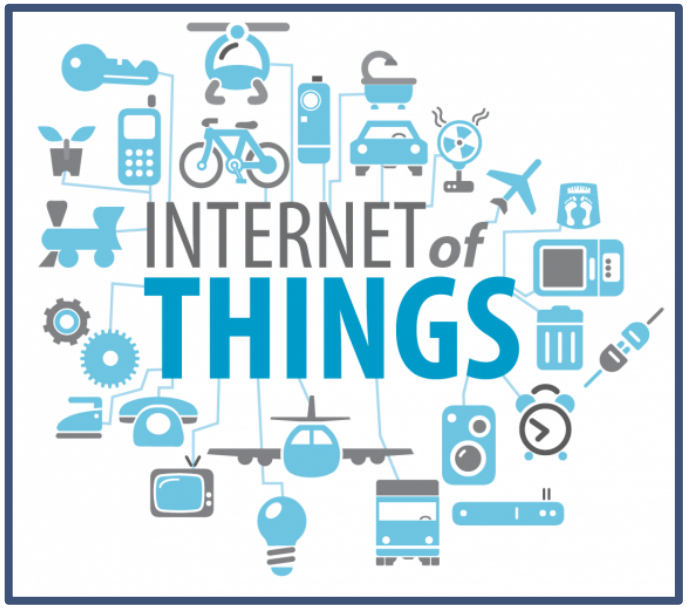
DR. MARDHANI RIASETIAWAN
CAROLUS GAZA NINDRA TAMA - 13/347460/PA/15250

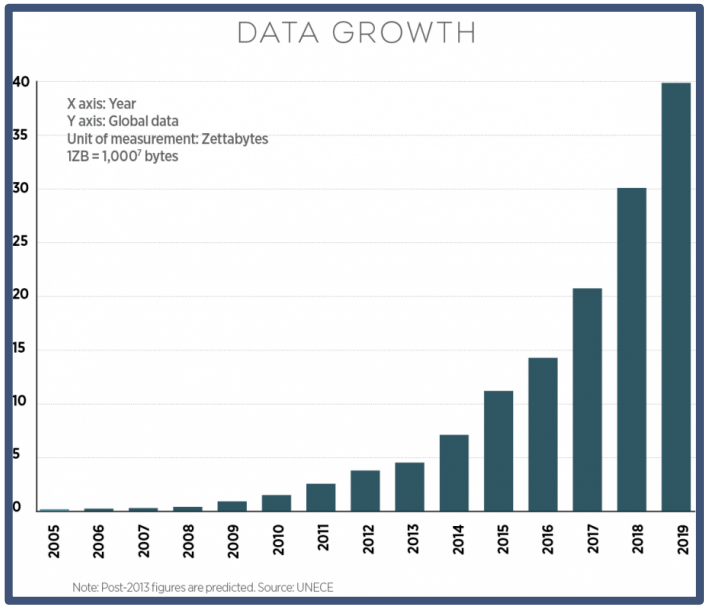
ILMU KOMPUTER
UNIVERSITAS GADJAH MADA

1

PENDAHULUAN

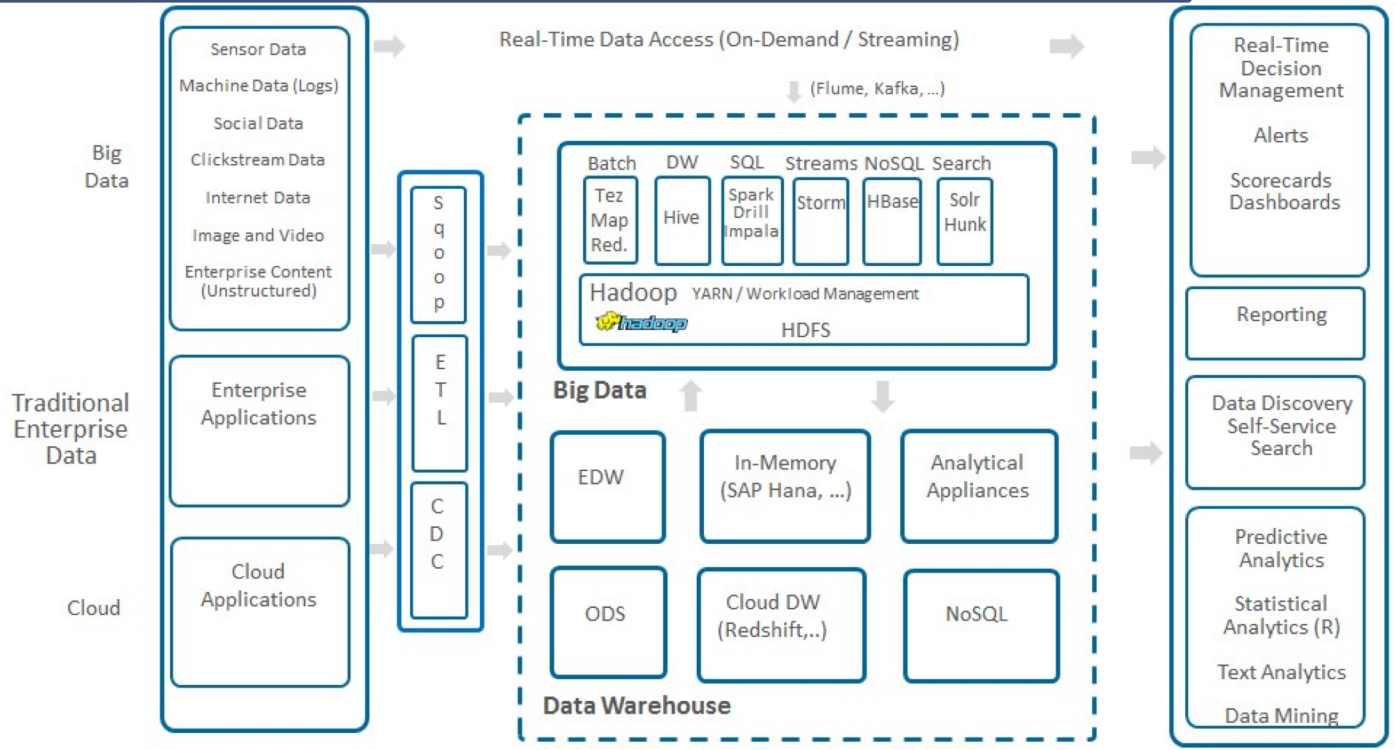
LATAR BELAKANG





LATAR BELAKANG

- Terpisah - pisah
- Rumit
- Menyita waktu


Big Data
Traditional Enterprise Data
Cloud
Sensor Data
Machine Data (Logs)
Social Data
Clickstream Data
Internet Data
Image and Video
Enterprise Content (Unstructured)
Enterprise Applications
Cloud Applications
Sqoop
ETL
CDC
Real-Time Data Access (On-Demand / Streaming)
(Flume, Kafka, ...)
Batch
DW
SQL
Streams
NoSQL
Search
Tez Map Red.
Hive
Spark Drill Impala
Storm
HBase
Solr Hunk
Hadoop
YARN / Workload Management
HDFS
Big Data
EDW
In-Memory (SAP Hana, ...)
Analytical Appliances
ODS
Cloud DW (Redshift,..)
NoSQL
Data Warehouse
Real-Time Decision Management
Alerts
Scorecards Dashboards
Reporting
Data Discovery Self-Service Search
Predictive Analytics
Statistical Analytics (R)
Text Analytics
Data Mining

RUMUSAN MASALAH

BAGAIMANA MERANCANG SISTEM OPERASI YANG SIAP UNTUK IMPLEMENTASI KLASTER DAN MEMILIKI MODUL UNTUK PENGOLAHAN BIG DATA DENGAN BERBASIS CENTOS 7.

BATASAN MASALAH

- Sistem operasi yang digunakan dalam penelitian ini adalah Centos 7
- Tool untuk klaster yang diimplementasikan adalah Message Passing Interface (MPI)
- Framework big data yang diimplementasikan adalah Apache Hadoop
- Tool untuk batch processing big data yang diimplementasikan adalah MapReduce
- Tool untuk stream processing big data yang diimplementasikan adalah Apache Spark

BATASAN MASALAH

- Data warehouse yang digunakan adalah Apache Hive dan Apache HBase
- Tool untuk scripting analysis yang diimplementasikan adalah Apache Pig
- Tool untuk koordinasi klaster yang diimplementasikan adalah Apache ZooKeeper
- Remastering yang dilakukan pada penelitian ini hanya terbatas pada modifikasi paket aplikasi dengan menggunakan kickstart

TUJUAN PENELITIAN

TUJUAN DARI PENELITIAN INI ADALAH MENGHASILKAN SUATU SISTEM OPERASI YANG MEMILIKI KEMAMPUAN UNTUK KLASTER DAN EKOSISTEM BIG DATA YANG DAPAT DIGUNAKAN UNTUK MELAKUKAN PEMROSESAN BIG DATA.

MANFAAT PENELITIAN

MENYEDIAKAN SUATU SISTEM OPERASI YANG SUDAH SIAP DIIMPLEMENTASIKAN PADA KOMPUTER KLASTER DAN MEMILIKI BEBERAPA FITUR YANG TELAH MENDUKUNG UNTUK PENGOLAHAN BIG DATA SEHINGGA MEMUDAHKAN PENGGUNA DALAM MEMBUAT SUATU EKOSISTEM BIG DATA DALAM KOMPUTER KLASTER.

2

TINJAUAN PUSTAKA

	Peneliti	Judul	Penelitian
1	Santosa dkk (2010)	*Remastering Distro Ubuntu untuk menunjang Pembelajaran Informatika*	Membuat sistem operasi linux yang bertujuan untuk menunjang pembelajaran informatika. Metode yang digunakan adalah remastering dari sistem operasi Ubuntu 9.04 (*Jaunty Jackalope*)
2	Sulistyo, A. (2010)	*Membuat Distro Linux untuk Security*	Membuat sistem operasi linux yang bertujuan untuk mencegah serangan eksploitasi yang memanfaatkan bug pada kernel. Metode yang digunakan adalah remastering dari sistem operasi Slax 6.0.9.
3	Mazumdar dan Dhar (2015)	*Hadoop as Big Data Operating System – The Emerging Approach for Managing Challenges of Enterprise Big Data Platform*	Memberikan pendekatan untuk mengatasi tantangan big data dengan menggunakan framework Hadoop.
4	Ji, C. dkk (2015)	*IBDP: An Industrial Big Data Ingestion and Analysis Platform and Case Studies*	Membangun Big Data Platform untuk keperluan dalam bidang industri.

	Peneliti	Judul	Penelitian
5	Liu, L. (2015)	*Performance Comparison by Running Benchmarks on Hadoop, Spark, and HAMR*	Melakukan perbandingan performa pada platform bigdata Hadoop, Spark, dan HAMR.
6	Pan, S. (2016)	*The Performance Comparison of Hadoop and Spark*	Melakukan perbandingan performa pada Hadoop dan Spark dengan menggunakan benchmarks WordCount, Sort, dan PageRank.

3

DESAIN DAN IMPLEMENTASI

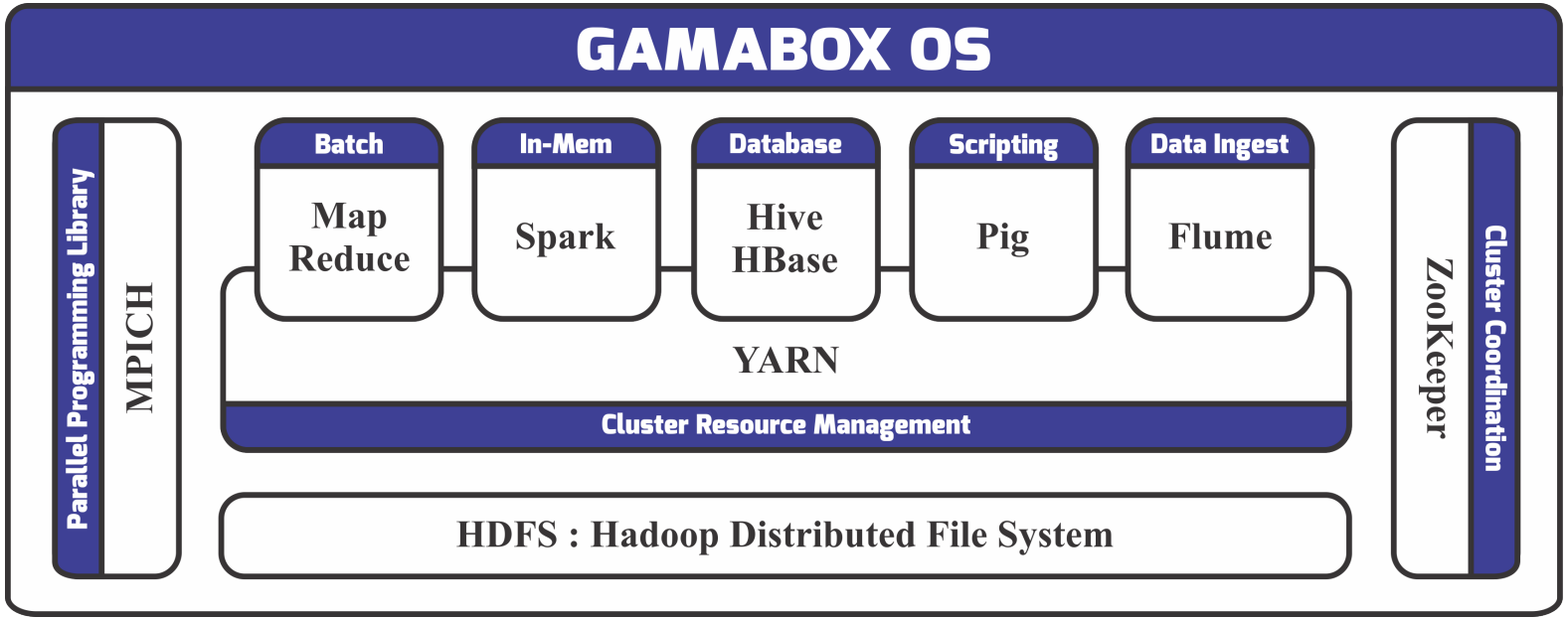
SPESIFIKASI PERANGKAT KERAS

	Perangkat Keras	Nama
1	*Processor*	Intel(R) Core(TM) i3-4170 CPU @ 3.70GHz
	Storage	HDD 1TB
	RAM	12 GB
2	*Processor*	Intel(R) Core(TM) i3-4170 CPU @ 3.70GHz
	Storage	HDD 500GB
	RAM	12 GB

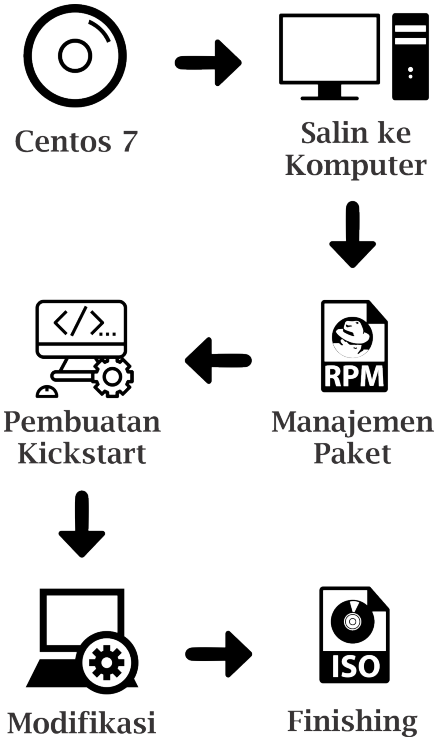
SPESIFIKASI PERANGKAT LUNAK

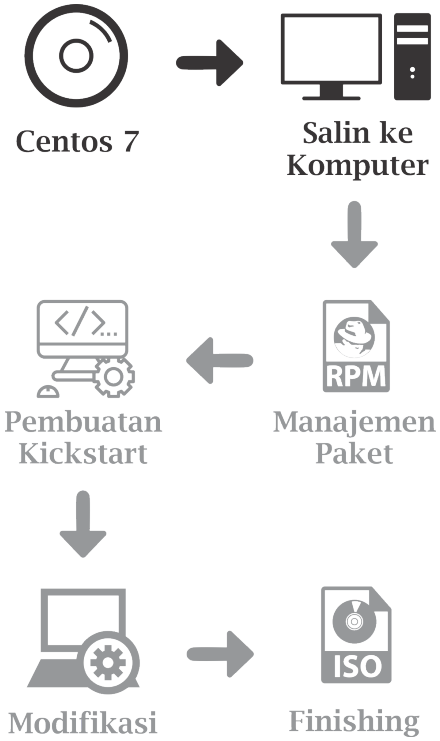
	Perangkat Lunak	Nama
1	*Sistem Operasi*	Centos 7
2	*Cluster*	Apache Hadoop
3	*Data Management*	HDFS
4	*Data Processing*	MapReduce, Spark
5	*Database*	Hive, Pig, HBase
6	*Machine Learning*	Spark MLlib
7	*Data Ingestion*	Apache Flume

RANCANGAN ARSITEKTUR SISTEM OPERASI



PROSES REMASTERING SISTEM OPERASI

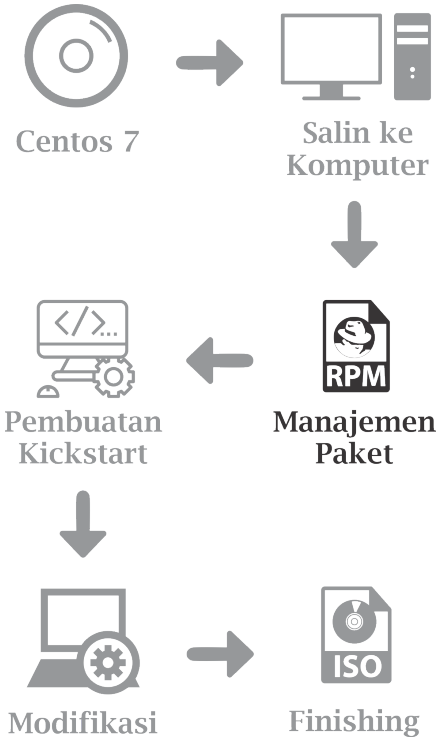





```
$ sudo mount -o loop -t iso9660 Centos7.iso
  /mnt/
```

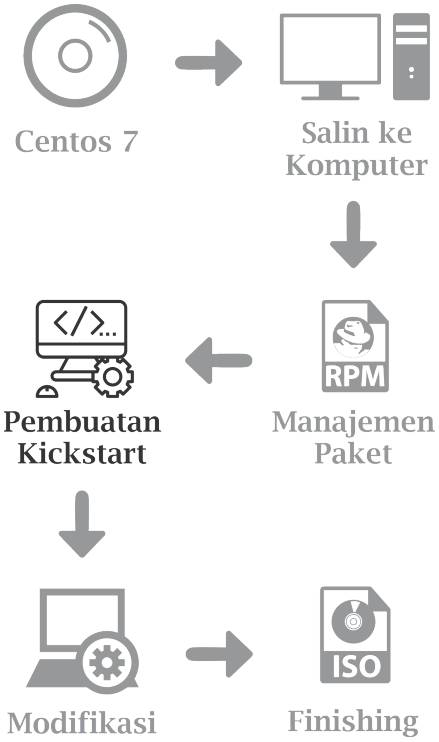
```
$ rysnc -av /mnt/ /home/labskj/remaster/
```

```
$ sudo find ./ -name TRANS.TBL -exec rm -f {}
  \; -print
```

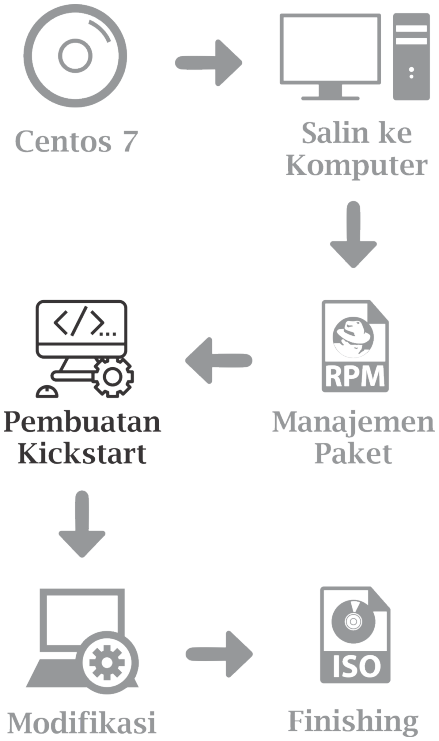



```
$ ./gather_packages.pl /mnt/repodata/*comps*.xml
  /mnt/Packages/ /home/labskj/remaster/Packages/
  x86_64 base core debugging development compat-
  libraries fonts gnome-desktop internet-browser
  network-file-system-client postgresql postgresql-
  client scientific x11 python-devel
```

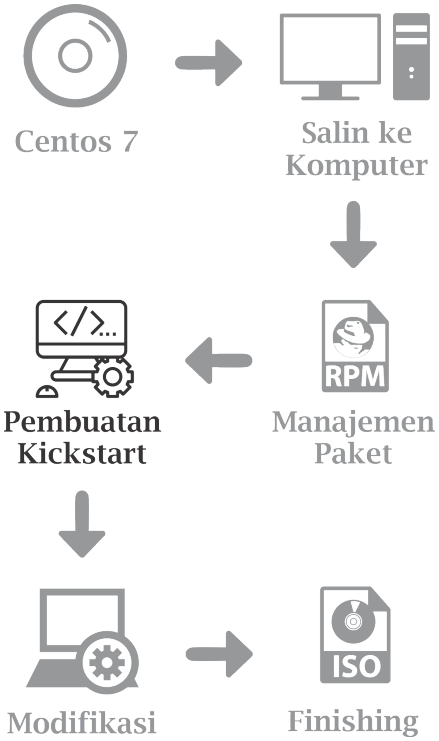
```
$ ./resolve_deps.pl /mnt/Packages/
  /home/labskj/remaster/Packages/ x86_64
```



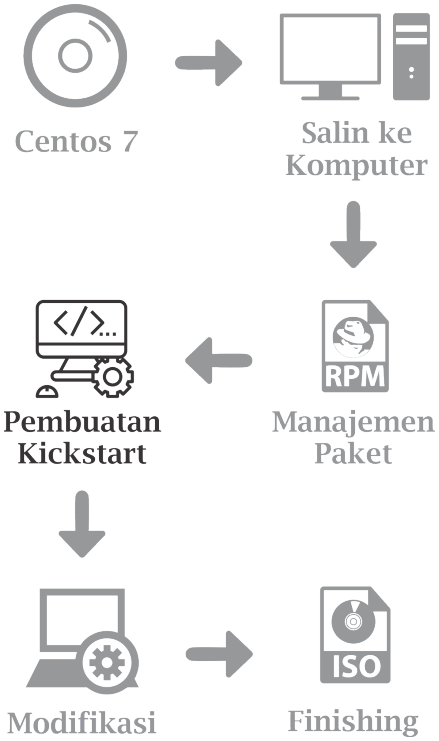

```
1.    xconfig  --startxonboot
2.    eula -agreed
3.    auth --enableshadow -passalgo=sha512
4.    cdrom
5.    graphical
6.    firstboot --disable
7.    keyboard --vckeymap=us --xlayouts='us'
8.    lang en_US.UTF-8
9.    ignoredisk -only-use=sda
10.   network  --bootproto=dhcp --device=enp2s0 --onboot=off --ipv6=auto
11.   network  --bootproto=dhcp --hostname=localhost.localdomain
12.   firewall --disable
13.   services --enabled=NetworkManager,sshd
14.   rootpw --iscrypted
      $6$CXp7SwZ/O9mm2qbx$z/0reUhawrGI/eXt.qOgQMMzwT0KtGj9P/M4rFtYpEByRtBjGmXZ
      26EVJWt4CRnBK54LJTH.2BAWE9ADo3wEX.
15.   timezone America/New_York -isUtc
16.   group -name=hadoop
17.   user --name=hduser --groups=hadoop --
      password=$6$btBNMpnUqfWu68He$pqlwQsNpAwbeXfxHhm5KUsmS31CTECINQo5ZtHnm8bF
      18sCnOkrNQ9NL1C/uNR6t1gr7gms6piXLFq9GDESHh/ --iscrypted --gecos="Hadoop
      User"
18.   clearpart --all --initlabel --drives=sda
19.   bootloader --append=" crashkernel=auto" --location=mbr --boot-drive=sda
20.   autopart --type=lvm
21.   reboot
```



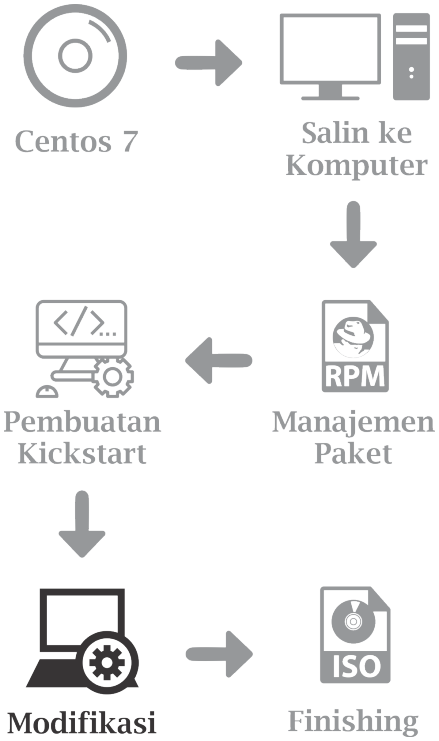

```
22.  %packages
23.  @base
24.  @core
25.  @debugging
26.  @development
27.  @compat-libraries
28.  @fonts
29.  @gnome-desktop
30.  @internet-browser
31.  @network-file-system-client
32.  @postgresql
33.  @postgresql-client
34.  @scientific
35.  @x11
36.  kexec-tools
37.  python-devel
38.  %end
```



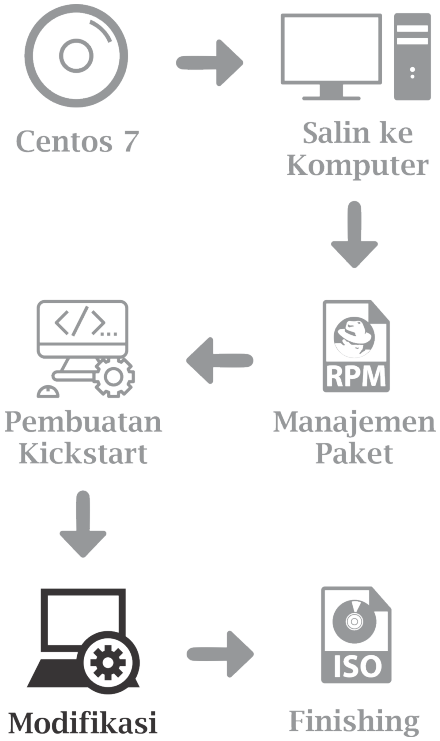

```
39.  %post --nochroot
40.  #!/bin/sh
41.
42.  # Copy all files to /root/ folder
43.  cp -r /run/install/repo/postinstall /mnt/sysimage/root
44.
45.  %end
```




```
46.  %post
47.  #!/bin/sh
48.
49.  /root/postinstall/script/install-common-tools.sh
50.  /root/postinstall/script/install-mpich.sh
51.  /root/postinstall/script/install-hadoop.sh
52.  /root/postinstall/script/install-zookeeper.sh
53.  /root/postinstall/script/install-pig.sh
54.  /root/postinstall/script/install-hive.sh
55.  /root/postinstall/script/install-hbase.sh
56.  /root/postinstall/script/install-flume.sh
57.  /root/postinstall/script/install-spark.sh
58.
59.  # Cleaning installation files
60.  rm -rf /root/*
61.  rm -rf /opt/install-env.sh
62.
63.  %end
```



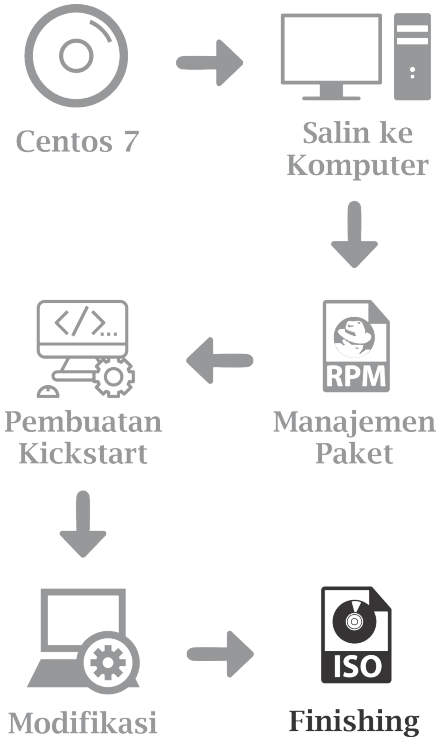

```
1.    menu title Gamabox OS
2.
3.    ...
4.
5.    label linux
6.      menu label ^Install Gamabox OS
7.      kernel vmlinuz
8.      append initrd=initrd.img
      inst.stage2=hd:LABEL=CentOS\x207\x20x86_64
      inst.ks=file:/ks.cfg
```




```
$ mv initrd.img initrd.img.xz
$ mkdir irmod
$ cd irmod
$ unxz ../initrd.img.xz
$ cpio -id < ../initrd.img
```

```
$ cp ../../ks/ks.cfg .
```

```
$ find . | cpio -o -H newc > ../new.img
$ cd ..
$ xz -C crc32 new.img
$ mv new.img.xz initrd.img
$ rm -rf irmod
```




```
$ cd /home/labskj/remaster/
$ createrepo -g /mnt/repodata/*comps*.xml .
```

```
$ cd /home/labskj/
$ chmod 664 remaster/isolinux/isolinux.bin
$ mkisofs -o GamaboxOS.iso -b isolinux.bin -c boot.cat \
  -no-emul-boot \
  -V 'CentOS 7 x86_64' \
  -boot-load-size 4 \
  -boot-info-table -R -J -v -T remaster/isolinux/
```

PENGUJIAN

- FUNGSIONALITAS
- PERFORMA

PENGUJIAN FUNGSIONALITAS

- ✓ Sistem Operasi
- ✓ HDFS
- ✓ Hadoop MapReduce
- ✓ Apache Spark
- ✓ Apache Flume
- ✓ Apache Hive
- ✓ Apache Pig
- ✓ Apache ZooKeeper
- ✓ Apache HBase

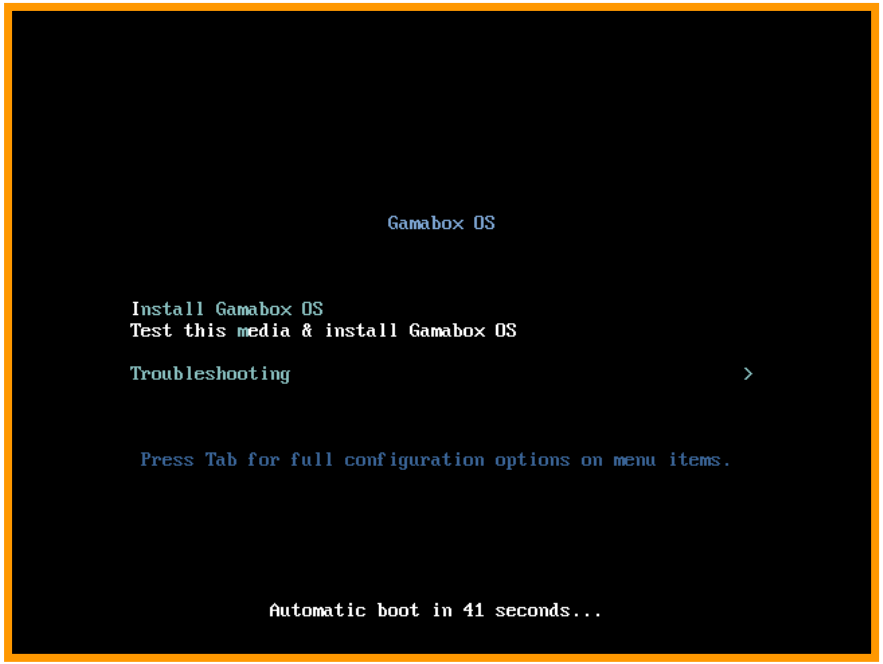
PENGUJIAN PERFORMA

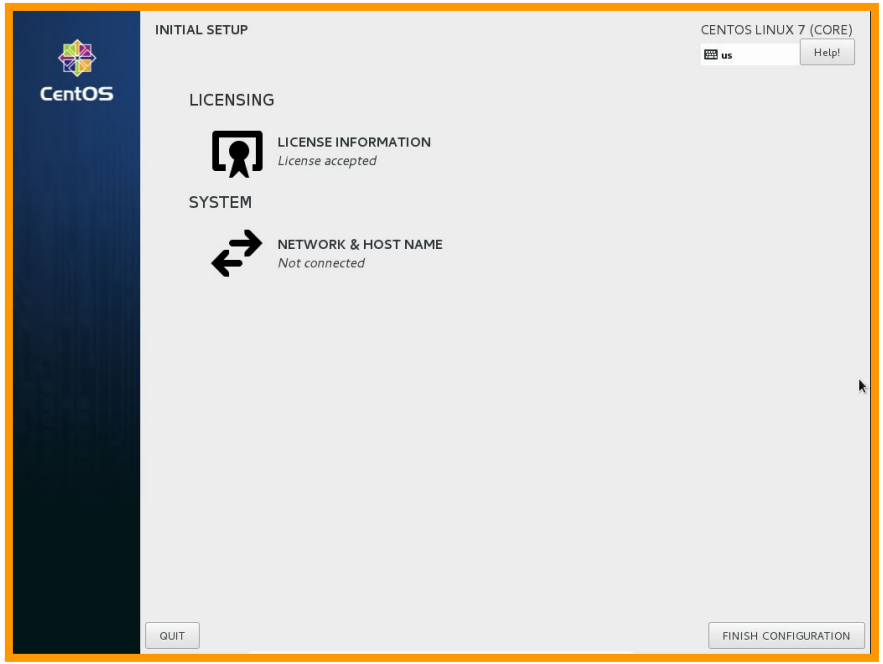
Membandingkan performa Gamabox OS dengan HDP:

- Pemantauan Kinerja (Memory dan CPU)
- Benchmark TestDFSIO
- Benchmark TeraSort

4

HASIL PENGUJIAN


Gamabox OS
Install Gamabox OS
Test this media & install Gamabox OS
Troubleshooting
Press Tab for full configuration options on menu items.
Automatic boot in 41 seconds...


CentOS
INITIAL SETUP
CENTOS LINUX 7 (CORE)
us
Help!
LICENSING
LICENSE INFORMATION
License accepted
SYSTEM
NETWORK & HOST NAME
Not connected
QUIT
FINISH CONFIGURATION

```
[hduser@10 ~]$ hadoop fs -put coba .
[hduser@10 ~]$ hadoop fs -ls
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2017-04-21 15:57 .sparkStaging
-rw-r--r--   1 hduser supergroup         44 2017-05-11 14:24 coba
```

```
[hduser@10 ~]$ hadoop fs -cat coba
Carolus Gaza Nindra Tama
Ilmu Komputer UGM
```

```
[hduser@10 coba]$ ls -l
total 0
[hduser@10 coba]$ hadoop fs -get coba .
[hduser@10 coba]$ ls -l
total 4
-rw-r--r--. 1 hduser hduser 44 May 11 14:32 coba
```

```
[hduser@10 mapreduce]$ hadoop jar hadoop-mapreduce-examples-2.7.3.jar wordcount coba coba.out
17/05/11 15:06:33 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/05/11 15:06:34 INFO input.FileInputFormat: Total input paths to process : 1
17/05/11 15:06:34 INFO mapreduce.JobSubmitter: number of splits:1
17/05/11 15:06:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1494336581947_0002
17/05/11 15:06:35 INFO impl.YarnClientImpl: Submitted application application_1494336581947_0002
17/05/11 15:06:35 INFO mapreduce.Job: The url to track the job: http://10.6.252.8:8088/proxy/applica
tion_1494336581947_0002/
17/05/11 15:06:35 INFO mapreduce.Job: Running job: job_1494336581947_0002
17/05/11 15:06:40 INFO mapreduce.Job: Job job_1494336581947_0002 running in uber mode : false
17/05/11 15:06:40 INFO mapreduce.Job:  map 0% reduce 0%
17/05/11 15:06:44 INFO mapreduce.Job:  map 100% reduce 0%
17/05/11 15:06:48 INFO mapreduce.Job:  map 100% reduce 100%
17/05/11 15:06:49 INFO mapreduce.Job: Job job_1494336581947_0002 completed successfully
```

```
[hduser@10 mapreduce]$ hadoop fs -cat coba.out/part-r-00000
Carolus 1
Gaza    1
Ilmu    1
Komputer        1
Nindra  1
Tama    1
UGM     1
```

```
scala> var Data = sc.textFile("/home/hduser/coba/coba")
Data: org.apache.spark.rdd.RDD[String] = /home/hduser/coba/coba MapPartitionsR
DD[1] at textFile at <console>:24

scala> var tokens = Data.flatMap(s => s.split(" "))
tokens: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <
console>:26

scala> var tokens_1 = tokens.map(s => (s,1))
tokens_1: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map
 at <console>:28

scala> var sum_each = tokens_1.reduceByKey((a, b) => a + b)
sum_each: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceBy
Key at <console>:30

scala> sum_each.collect()
res0: Array[(String, Int)] = Array((Nindra,1), ("",1), (Gaza,1), (Tama,1), (Ca
rolus,1), (Komputer,1), (UGM,1), (Ilmu,1))

scala> sum_each.saveAsTextFile("/home/hduser/coba/spark_out")
```

```
[hduser@10 ~]$ head /home/hduser/coba/spark_out/part-00000
(Nindra,1)
(,1)
[hduser@10 ~]$ head /home/hduser/coba/spark_out/part-00001
(Gaza,1)
(Tama,1)
(Carolus,1)
(Komputer,1)
(UGM,1)
(Ilmu,1)
```

Browse Directory

/user/flume/tweets/2017/04/11/10 Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	17.83 KB	5/16/2017, 3:00:13 PM	1	128 MB	FlumeData.1491880605891
-rw-r--r--	hduser	supergroup	23.28 KB	5/16/2017, 3:00:13 PM	1	128 MB	FlumeData.1491880605892
-rw-r--r--	hduser	supergroup	26.08 KB	5/16/2017, 3:00:13 PM	1	128 MB	FlumeData.1491880605893
-rw-r--r--	hduser	supergroup	37.8 KB	5/16/2017, 3:00:14 PM	1	128 MB	FlumeData.1491880605894
-rw-r--r--	hduser	supergroup	8.01 KB	5/16/2017, 3:00:14 PM	1	128 MB	FlumeData.1491880605895

```
hive> LOAD DATA LOCAL INPATH '/home/hduser/Documents/rawTweets/*/*/*/*/*' INTO TABLE tweets;
Loading data to table twitter.tweets
OK
Time taken: 68.591 seconds
```

```
2017-05-14 21:01:57,797 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 245.1 sec
MapReduce Total cumulative CPU time: 4 minutes 5 seconds 100 msec
Ended Job = job_1494748546030_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 8  Reduce: 1   Cumulative CPU: 245.1 sec   HDFS Read: 1854777339 HDFS Write: 106
 SUCCESS
Total MapReduce CPU Time Spent: 4 minutes 5 seconds 100 msec
OK
335890
Time taken: 104.331 seconds, Fetched: 1 row(s)
```

```
2017-05-16 15:24:16,371 [main] INFO  org.apache.pig.backend.hadoop.executioneng
ine.util.MapRedUtil - Total input paths to process : 1
(1,Carolus,Gaza,21,85729058258,Klaten)
(5,Yoga,Raharja,21,87585646585,Surabaya)
(3,Andreas,Dimas,22,87469852123,Solo)
(6,Maulana,Kamil,23,87456852468,Jakarta )
2017-05-16 15:24:16,402 [main] INFO  org.apache.pig.Main - Pig script completed
 in 4 seconds and 384 milliseconds (4384 ms)
```

```
[hduser@10 ~]$ zkServer.sh start
ZooKeeper JMX enabled by default
Using config: /opt/zookeeper-3.4.9/bin/../conf/zoo.cfg
Starting zookeeper ... STARTED
[hduser@10 ~]$ jps
30582 QuorumPeerMain
30604 Jps
```

```
WATCHER::

WatchedEvent state:SyncConnected type:None path:null

[zk: localhost:2181(CONNECTED) 0]
```

```
hbase(main):001:0> create 'table','test'
0 row(s) in 2.5240 seconds

=> Hbase::Table - table
hbase(main):002:0> list
TABLE
table
1 row(s) in 0.0220 seconds

=> ["table"]
```

```
hbase(main):004:0> disable 'table'
0 row(s) in 2.2760 seconds

hbase(main):005:0> drop 'table'
0 row(s) in 1.2680 seconds

hbase(main):006:0> list
TABLE
0 row(s) in 0.0040 seconds

=> []
```

PENGUJIAN FUNGSIONALITAS

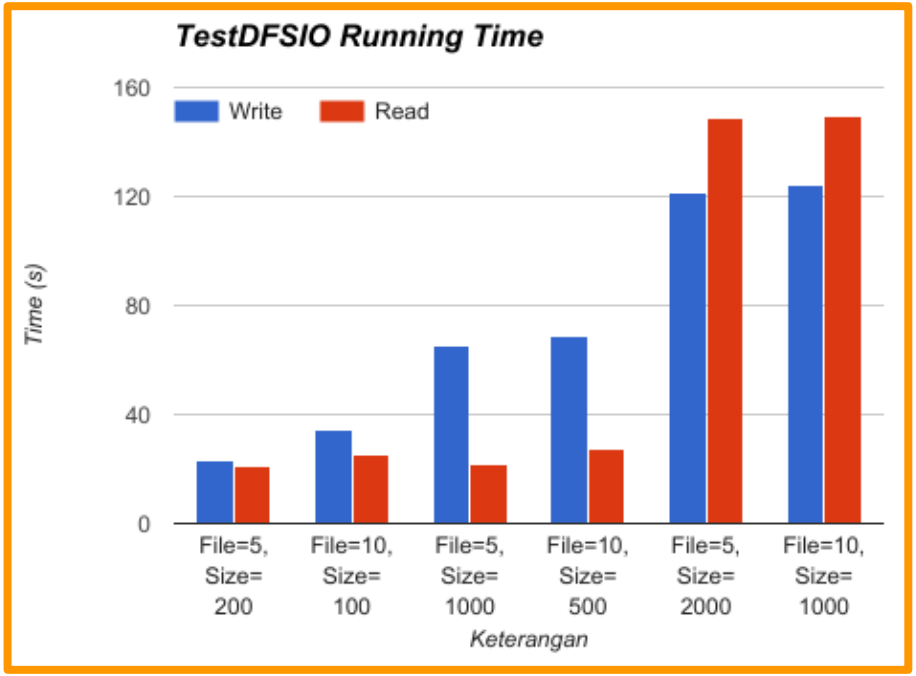
	Pengujian	Status
1	*Instalasi Sistem Operasi*	Berhasil
2	*HDFS*	Berhasil
3	*Hadoop MapReduce*	Berhasil
4	*Apache Spark*	Berhasil
5	*Apache Flume*	Berhasil
6	*Apache Hive*	Berhasil
7	*Apache Pig*	Berhasil
8	*Apache ZooKeeper*	Berhasil
9	*Apache HBase*	Berhasil

Pemantauan Kinerja Gamabox OS

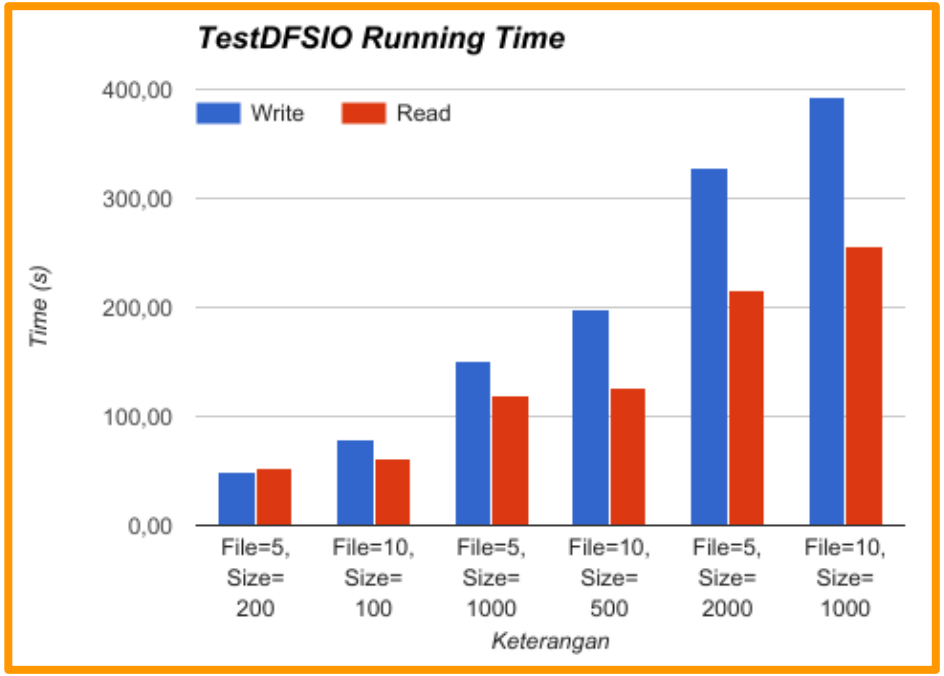
	Status	Memory	CPU
1	*Sistem*	2,69%	0,25%
2	*Hadoop*	11,94%	1,08%
3	*Hadoop + Spark*	15,70%	1,34%
4	*Hadoop + Spark + HBase*	19,09%	1,23%
5	*Hadoop + Spark + HBase + Hive*	22,44%	2,91%
6	*Hadoop + Spark + HBase + Hive + Pig (semua services)*	24,18%	1,34%

Pemantauan Kinerja HDP

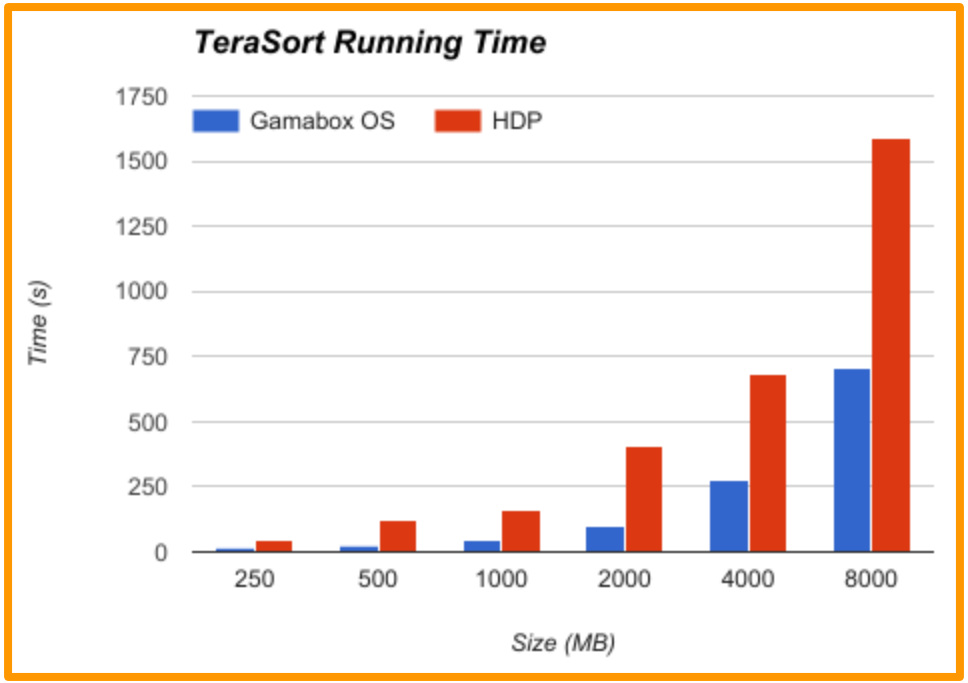
	Status	Memory	CPU
1	*Sistem*	7,66%	10,09%
2	*Ambari*	12,26%	10,63%
3	*Semua services*	69,24%	21,98%

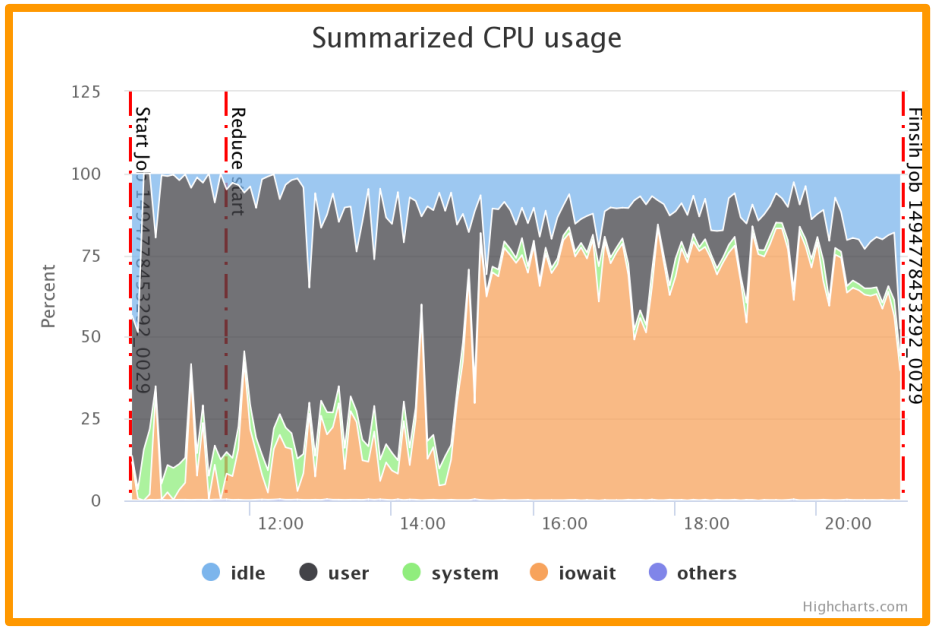


Gamabox OS

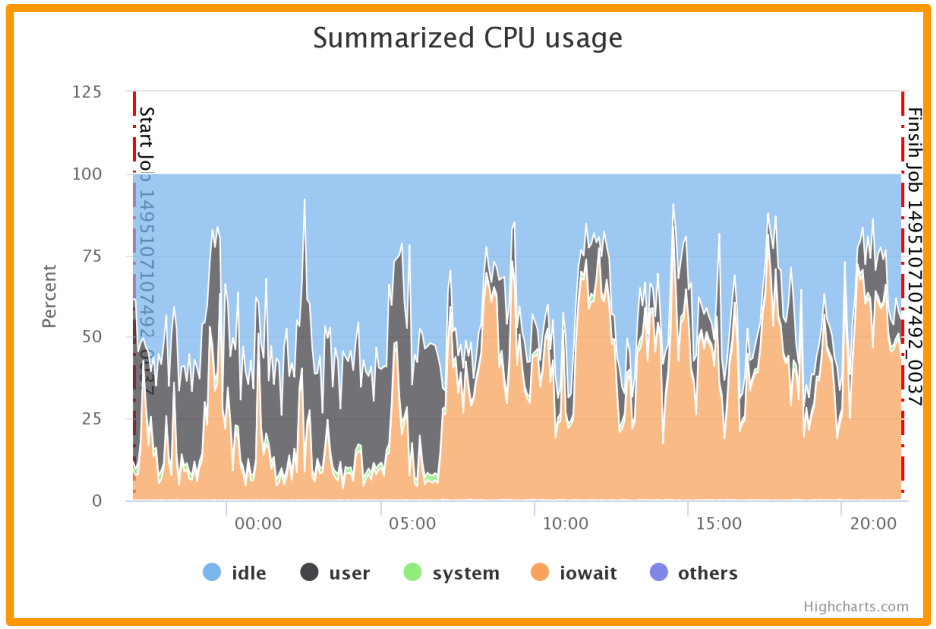


HDP


TeraSort Running Time
1750
1500
1250
1000
750
500
250
0
Time (s)
Gamabox OS
HDP
250
500
1000
2000
4000
8000
Size (MB)



Gamabox OS



HDP

5

KESIMPULAN DAN SARAN

KESIMPULAN

1. Aplikasi Hadoop, Spark, Flume, Hive, Pig, HBase, dan ZooKeeper yang diimplementasikan pada Gamabox OS dapat berjalan dengan lancar.
2. Sistem operasi Gamabox OS memiliki penggunaan memori dan CPU yang lebih baik dibandingkan dengan HDP. Hal ini ditunjukkan dengan menjalankan semua service pada kondisi idle, Gamabox OS hanya menggunakan memori sebesar 24,8% dan penggunaan CPU sebesar 1,34%.

KESIMPULAN

3. Sistem operasi Gamabox OS memiliki performa yang lebih baik daripada HDP. Tetapi masih tidak stabil ketika melakukan pengujian baca data pada TestDFSIO.
4. Penggunaan CPU pada sistem operasi Gamabox OS saat menjalankan benchmark sudah bekerja dengan baik, tetapi masih belum optimal.

SARAN

1. Diperlukan penelitian lebih lanjut mengenai optimalisasi performa pada masing-masing aplikasi big data.
2. Diperlukan penelitian untuk menggabungkan komputer-komputer dengan sistem operasi Gamabox OS menjadi suatu komputer klaster.

THANK YOU!