2021

Big Data / Lab04

TEAM KAPPA

WEIDELE, PRÜLL, TOMONDY, BAUER

HDInsight cluster

HDInsight-Cluster erstellen ...

Grundlagen | Speicher | Sicherheit + Netzwerkbetrieb | Konfiguration + Preise | Tags | Bewerten + erstellen

Neu bei HDInsight? Starten Sie mit unseren Schulungsressourcen..
Erstellen Sie einen verwalteten HDInsight-Cluster. Treffen Sie Ihre Auswahl zwischen Spark, Kafka, Hadoop, Storm und vielen anderen. Weitere Informationen

Projektdetails

Wählen Sie das Abonnement aus, um bereitgestellte Ressourcen und Kosten zu verwalten. Verwenden Sie Ressourcengruppen wie z. B. Ordner zum Organisieren und Verwalten all Ihrer Ressourcen.

Abonnement * : Azure für Bildungseinrichtungen

Ressourcengruppe * : Meine-Ressourcengruppe

Neues Element erstellen

Clusterdetails

Benennen Sie Ihren Cluster, wählen Sie eine Region sowie einen Clustertyp und eine Clusterversion aus. Weitere Informationen

Clustername * : Lab04-Cluster

Region * : Deutschland, Westen-Mitte

Clustertyp * : **Spark**
Änderung

Version * : Spark 2.4 (HDI 4.0)

Clusteranmeldeinformationen

Geben Sie neue Anmeldeinformationen ein, die zur Verwaltung oder für den Zugriff auf den Cluster verwendet werden.

Benutzername für Clusteranmeldung * : admin

Kennwort für Clusteranmeldung * : ••••••••••

Clusteranmeldekennwort bestätigen * : ••••••••••

SSH-Benutzername (Secure Shell) * : sshuser

Verwenden Sie ein Clusteranmeldekennwort für SSH. : ☑

Grundlagen Speicher Sicherheit + Netzwerkbetrieb **Konfiguration + Preise** Tags Bewerten + erstellen

Konfigurieren Sie Clusterleistung und Preise. Weitere Informationen

Knotenkonfiguration

Konfigurieren Sie Größe und Leistung Ihres Clusters, und zeigen Sie Informationen zu den geschätzten Kosten an.

Die Kostenschätzung in der Tabelle berücksichtigt keine Abonnementrabatte oder Kosten in Zusammenhang mit Speicher, Netzwerk oder Datenübertragung.

Diese Konfiguration verwendet 38 von 40 verfügbaren Kernen in der Region "Deutschland, Westen-

Nutzung von Kernen anzeigen

+ Anwendung hinzufügen

Knotentyp	Knotengröße	Knotenanzahl	Geschätzte Koste...
Hauptknoten-K...	E8 V3 (8 Kerne, 64 GB RAM), 0.64 EUR/Stunde	2	1.28 EUR
Zookeeper-Knot...	A2 v2 (2 Kerne, 4 GB RAM), 0.10 EUR/Stunde	3	0.00 (KOSTENLOS)
Worker-Knoten	E8 V3 (8 Kerne, 64 GB RAM), 0.64 EUR/Stunde	2	1.28 EUR

☐ Automatische Skalierung aktivieren Weitere Informationen

Geschätzte Gesamtkosten/Stunde 2.56 EUR

Skriptaktionen

Verwenden Sie Skriptaktionen zum Ausführen benutzerdefinierter PowerShell- oder Bash-Skripts auf Clusterknoten während der Clusterbereitstellung. Informationen zu Skriptaktionen

+ Skriptaktion hinzufügen

HDInsight-Cluster erstellen ...

Überprüfung erfolgreich.

Grundlagen | Speicher | Sicherheit + Netzwerkbetrieb | Konfiguration + Preise | Tags | **Bewerten + erstellen**

Spark 2.4 (HDI 4.0)

2.56 EUR Geschätzte Gesamtkosten/Stunde
Diese Schätzung berücksichtigt keine Abonnementrabatte oder Kosten im Zusammenhang mit Speicher, Netzwerk oder Datenübertragung.

Grundlagen

Abonnement	Azure für Bildungseinrichtungen
Ressourcengruppe	Meine-Ressourcengruppe
Region	Deutschland, Westen-Mitte
Clustername	(neu) Lab04-Cluster
Clustertyp	Spark 2.4 (HDI 4.0)
Benutzername für Clusteranmeldung	admin
SSH-Benutzername (Secure Shell)	sshuser
Verwenden Sie ein Clusteranmeldekennwort für SSH.	Aktiviert

Sicherheit + Netzwerkbetrieb

TLS-Mindestversion	1.2
Ressourcenanbieterverbindung	Inbound
Verschlüsselung im Ruhezustand	Deaktiviert
Verschlüsselung während der Übertragung	Deaktiviert
Verschlüsselung auf dem Host für temporären Datenträger	Deaktiviert

Speicher

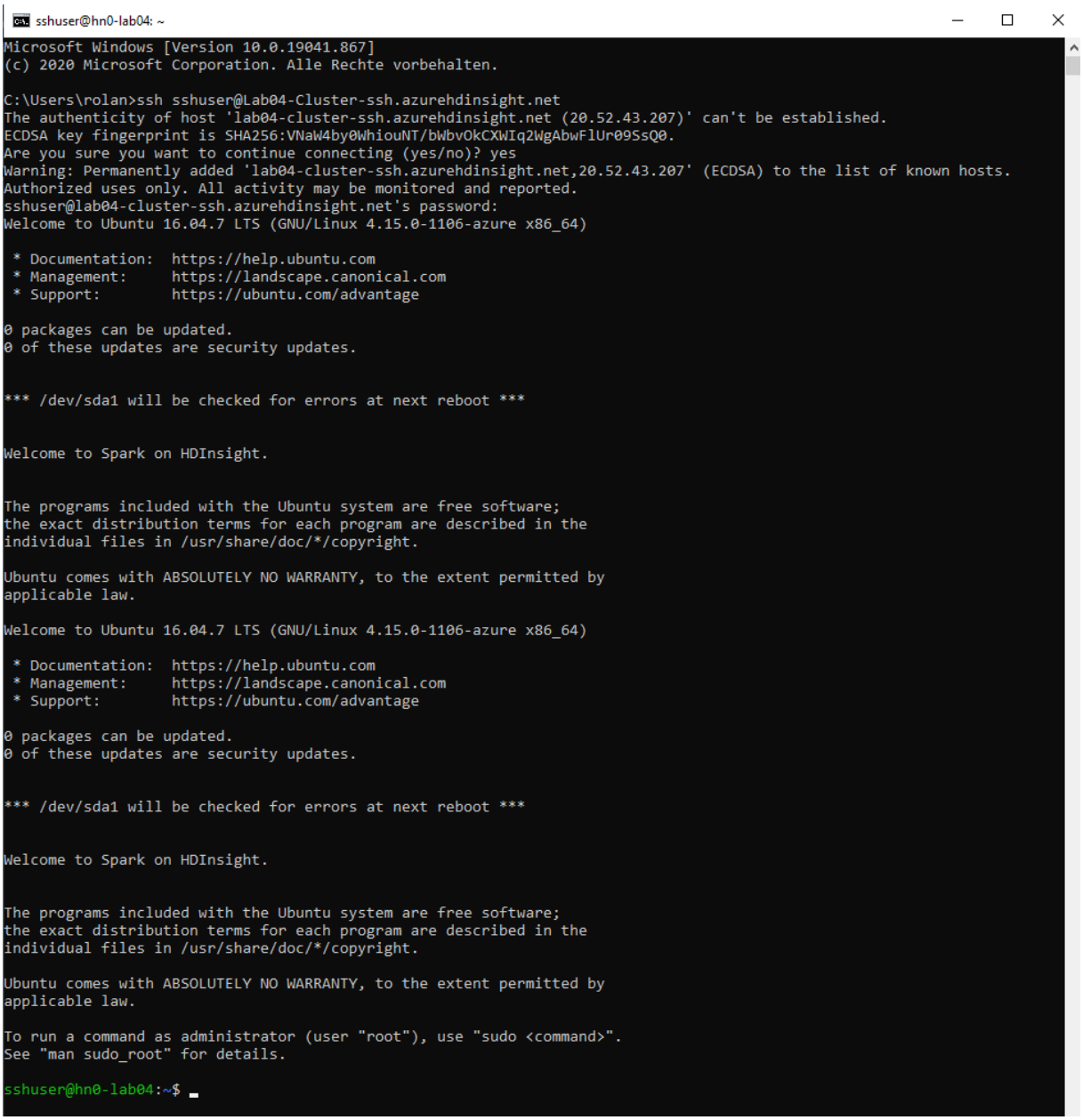
Primärer Speichertyp	Azure Storage
Primäres Speicherkonto	(neu) lab04storage
Container	lab04-cluster-2021-03-28t12-40-21-852z
Zusätzliche Azure Storage-Instanzen	Keine
Data Lake Storage Gen1-Zugriff	Deaktiviert

Clusterkonfiguration

Hauptknoten	2 Knoten, E8 V3 (8 Kerne, 64 GB RAM)
Zookeeper	3 Knoten, A2 v2 (2 Kerne, 4 GB RAM)
Worker	2 Knoten, E8 V3 (8 Kerne, 64 GB RAM)

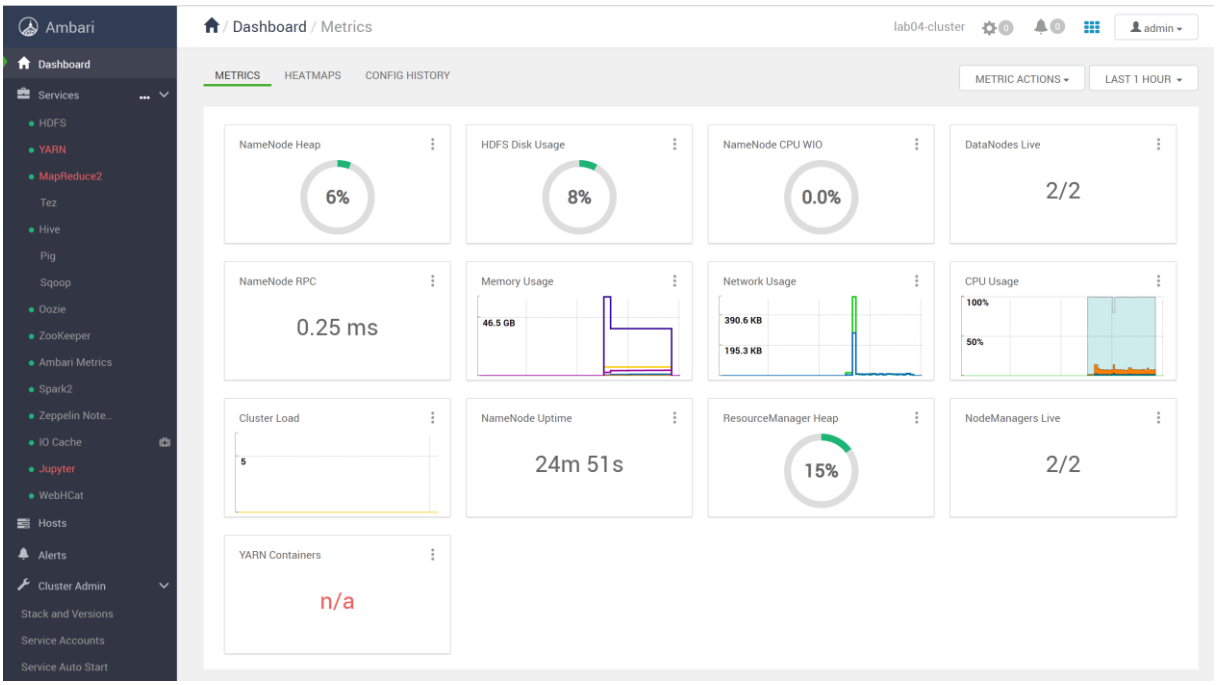
Erstellen | « Zurück | Weiter | Vorlage zur Automatisierung herunterladen

4. Explore the cluster



HDFS

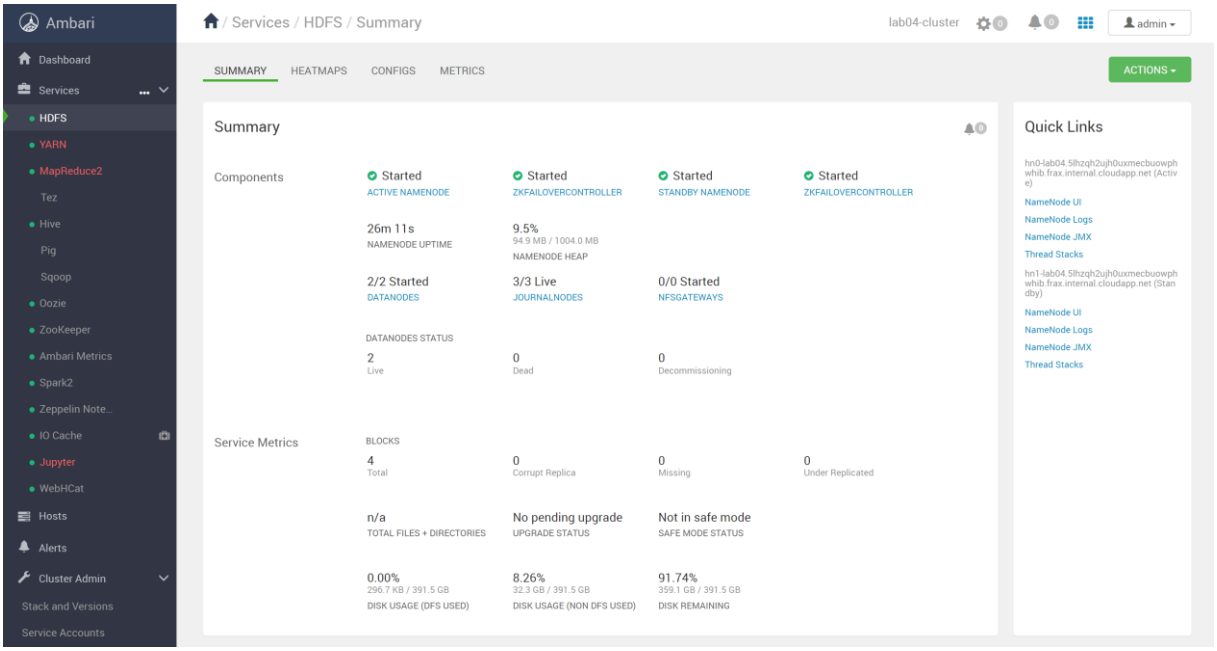
1. Login



2. Explore

hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net

hn1-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net



3. Connect

```
sshuser@hn0-lab04:~$ ssh hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net
The authenticity of host 'hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net (10.0.0.16)' can't be establis
hed.
ECDSA key fingerprint is SHA256:VNaW4by0WhiouNT/bWbvOkCXWIq2WgAbwFlUr09SsQ0.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net,10.0.0.16' (ECDSA) to the li
st of known hosts.
Authorized uses only. All activity may be monitored and reported.
sshuser@hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net's password:
Welcome to Ubuntu 16.04.7 LTS (GNU/Linux 4.15.0-1106-azure x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

0 packages can be updated.
0 of these updates are security updates.

New release '18.04.5 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** /dev/sda1 will be checked for errors at next reboot ***


Welcome to Spark on HDInsight.

Last login: Sun Mar 28 13:19:25 2021 from 89.144.206.17
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

sshuser@hn0-lab04:~$
```

4. HDFS NameNode

```
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

sshuser@hn0-lab04:~$ export HDFS=hdfs://hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudapp.net
sshuser@hn0-lab04:~$ hdfs dfs -mkdir $HDFS/lab04_rolands_dir
```

5. HDFS MKDIR

```
sshuser@hn0-lab04:~$ export HDFS=hdfs://hn0-lab04.5lhzqh2ujh0uxmecbuowphwhib.frax.internal.cloudap
sshuser@hn0-lab04:~$ hdfs dfs -mkdir $HDFS/lab04_rolands_dir
sshuser@hn0-lab04:~$ wget http://bit.ly/SeattleLibraryCheckoutRecords
```

6. DOWNLOAD FILES

```
sshuser@hn0-lab04:~$ hdfs dfs -mkdir $HDFS/lab04_folands_dir
sshuser@hn0-lab04:~$ wget http://bit.ly/SeattleLibraryCheckoutRecords
--2021-03-28 13:22:36--  http://bit.ly/SeattleLibraryCheckoutRecords
Resolving bit.ly (bit.ly)... 67.199.248.11, 67.199.248.10
Connecting to bit.ly (bit.ly)|67.199.248.11|:80... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: https://www.dropbox.com/s/x1n3olz6e5n7vqs/seattle-library-checkout-records.zip?dl=1 [following]
--2021-03-28 13:22:36--  https://www.dropbox.com/s/x1n3olz6e5n7vqs/seattle-library-checkout-records.zip?dl=1
Resolving www.dropbox.com (www.dropbox.com)... 162.125.65.18, 2620:100:6022:18::a27d:4212
Connecting to www.dropbox.com (www.dropbox.com)|162.125.65.18|:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: /s/dl/x1n3olz6e5n7vqs/seattle-library-checkout-records.zip [following]
--2021-03-28 13:22:36--  https://www.dropbox.com/s/dl/x1n3olz6e5n7vqs/seattle-library-checkout-records.zip
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://uc74df71633dae132c341793d29f.dl.dropboxusercontent.com/cd/0/get/BLhAJ5gnnVnwpx3l9HZfyJzyM8LSWTbvdbPy0y
kIn5zntvq8vIB-u2kTDWD4fI0PE8PW6aJlliz9jUoKmLUkXlmx_bSkumYFMkSsEBJGZXit7ozl95KIGIxZARrklX8qjDKAEAAKr0_ImTpK2RtsmdVw/file?
dl=1# [following]
--2021-03-28 13:22:37--  https://uc74df71633dae132c341793d29f.dl.dropboxusercontent.com/cd/0/get/BLhAJ5gnnVnwpx3l9HZfyJz
yM8LSWTbvdbPy0ykIn5zntvq8vIB-u2kTDWD4fI0PE8PW6aJlliz9jUoKmLUkXlmx_bSkumYFMkSsEBJGZXit7ozl95KIGIxZARrklX8qjDKAEAAKr0_ImTp
K2RtsmdVw/file?dl=1
Resolving uc74df71633dae132c341793d29f.dl.dropboxusercontent.com (uc74df71633dae132c341793d29f.dl.dropboxusercontent.com
)... 162.125.65.15, 2620:100:6022:15::a27d:420f
Connecting to uc74df71633dae132c341793d29f.dl.dropboxusercontent.com (uc74df71633dae132c341793d29f.dl.dropboxusercontent
.com)|162.125.65.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2511790568 (2.3G) [application/binary]
Saving to: 'SeattleLibraryCheckoutRecords'

SeattleLibraryCheckoutRecords 100%[=============================================>]   2.34G  22.3MB/s    in 2m 24s

2021-03-28 13:25:02 (16.7 MB/s) - 'SeattleLibraryCheckoutRecords' saved [2511790568/2511790568]

sshuser@hn0-lab04:~$ unzip SeattleLibraryCheckoutRecords -d records
Archive:  SeattleLibraryCheckoutRecords
  inflating: records/Checkouts_By_Title_Data_Lens_2008.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2015.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2013.csv
  inflating: records/Integrated_Library_System__ILS__Data_Dictionary.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2007.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2014.csv
  inflating: records/Library_Collection_Inventory.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2010.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2006.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2005.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2017.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2011.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2016.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2012.csv
  inflating: records/Checkouts_By_Title_Data_Lens_2009.csv
sshuser@hn0-lab04:~$
```

7. HDFS PUT

```
sshuser@hn0-lab04:~$ hdfs dfs -put Checkouts_By_Title_Data_Lens_2016.csv $HDFS/lab04dir/checkouts/
```

8. HDFS FSCK
    a. 4
    b. 1
    c. average block size: 115556230 Byte
        i. Block 1: 10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c
        ii. Block 2: 10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c
        iii. Block 3: 10.0.0.4:30010,DS-af7231e1-c0fc-4d70-aee1-276805a82901
        iv. Block 4: 10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c

```
sshuser@hn0-lab04:~$ hdfs fsck $HDFS/lab04dir/checkouts/ -files -blocks -locations
Connecting to namenode via http://hn0-lab04.xnl1ci4qj42evipvuy02v4heug.frax.internal.cloudapp.net:30070/fsck?ugi=sshuser&files=1&blocks=1&locations=1&path=%2Flab04dir%2Fcheckouts
FSCK started by sshuser (auth:SIMPLE) from /10.0.0.13 for path /lab04dir/checkouts at Mon Apr 05 18:28:11 UTC 2021
/lab04dir/checkouts <dir>
/lab04dir/checkouts/Checkouts_By_Title_Data_Lens_2016.csv 462224922 bytes, replicated: replication=1, 4 block(s):  OK
0. BP-1807273858-10.0.0.13-1617638491864:blk_1073743165_2341 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c,DISK]]
1. BP-1807273858-10.0.0.13-1617638491864:blk_1073743166_2342 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c,DISK]]
2. BP-1807273858-10.0.0.13-1617638491864:blk_1073743167_2343 len=134217728 Live_repl=1  [DatanodeInfoWithStorage[10.0.0.4:30010,DS-af7231e1-c0fc-4d70-aee1-276805a82901,DISK]]
3. BP-1807273858-10.0.0.13-1617638491864:blk_1073743168_2344 len=59571738 Live_repl=1  [DatanodeInfoWithStorage[10.0.0.6:30010,DS-df4e88cb-8382-4aca-8a4a-ce4d68c3e26c,DISK]]


Status: HEALTHY
 Number of data-nodes:  2
 Number of racks:               1
 Total dirs:                    1
 Total symlinks:                0

Replicated Blocks:
 Total size:    462224922 B
 Total files:   1
 Total blocks (validated):      4 (avg. block size 115556230 B)
 Minimally replicated blocks:   4 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Mon Apr 05 18:28:11 UTC 2021 in 2 milliseconds


The filesystem under path '/lab04dir/checkouts' is HEALTHY
sshuser@hn0-lab04:~$
```

9. PLAY AROUND WITH HDFS / query data with HIVE

HIVE

QUERY | JOBS | TABLES | SAVED QUERIES | UDFs | SETTINGS

create_table | Worksheet2 | +

DATABASE
Select or search database/schema
default

```
DROP TABLE lab04data;
CREATE EXTERNAL TABLE lab04data(
    BibNumber string,
    ItemBarcode string,
    ItemType string,
    BibCollection string,
    CallNumber string,
    CheckoutDateTime string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 'hdfs://hn0-lab04.y4tcsfbxfosevo2xcke2rpws2e.parx.internal.cloudapp.net/data/'
tblproperties ("skip.header.line.count"="1");

SELECT * FROM lab04data;
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter columns

lab04data.bibnumber	lab04data.itembarcode	lab04data.itemtype	lab04data.bibcollection	lab04data.callnumber	lab04data.checkoutdatetime
2054318	0010042526870	jcvhs	ncvid	VHS J BERENST	05/19/2005 11:18:00 AM
1630044	0010045312005	jcvhs	ncvidnf	VHS J597.3 Ey36S 1994	06/25/2005 03:11:00 PM
2203514	0010048118763	acbk	nanf	643.12 W4366C 2003	11/12/2005 04:42:00 PM
1988635	0010040727629	jcbk	ncpic	E BRADBUR	11/28/2005 04:38:00 PM
2119417	0010043384212	jcbk	ncpic	E WILLIAM	12/11/2005 02:52:00 PM
2273051	0010049785347	acdvd	nadvd	DVD AUTOBIO	07/29/2005 02:44:00 PM
2143833	0010045062352	acvhs	navid	FRENCH VHS THAT OB	09/07/2005 11:58:00 AM
2108178	0010045526679	acbk	nanf	745.582 St281B 2001	11/04/2005 05:55:00 PM

```
1 SELECT COUNT(*) AS TotalCount, lab04data.bibnumber, lab04inventory.title
2 FROM lab04data
3 JOIN lab04inventory ON (lab04data.bibnumber = lab04inventory.bibnumber)
4 GROUP BY lab04data.bibnumber, lab04inventory.title
5 ORDER BY totalcount DESC
6 LIMIT 10;
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter col

totalcount	lab04data.bibnumber	lab04inventory.title
841472	1923072	Guinness world records.
502911	1126205	The very hungry caterpillar / by Eric Carle.
494028	2525519	WALL-E [videorecording] / Walt Disney Pictures presents a Pixar Animation Studios film ; produced by Jim Morris ; original story & story by Andrew Stanton, Pete Docter ; screenplay by Andrew Stanton, Jim Reardon ; directed by Andrew Stanton.
452844	545653	Chicka chicka boom boom / by Bill Martin, Jr. and John Archambault ; illustrated by Lois Ehlert.
429640	2469502	Into the wild [videorecording] / Paramount Vantage ; River Road Entertainment ; a Square One C.I.H./Linson Film production ; produced by Sean Penn, Art Linson, Bill Pohlad ; screenplay and directed by Sean Penn.
412236	2277368	Harry Potter and the half-blood prince / by J.K. Rowling ; illustrations by Mary GrandPré.
408110	2609162	Up [videorecording] / Walt Disney Pictures ; a Pixar Animation Studios film ; produced by Jonas Rivera ; story by Pete Docter, Tom McCarthy, Bob Peterson ; screenplay by Bob Peterson, Pete Docter ; directed by Pete Docter, Bob Peterson.
396117	2124013	I stink! / Kate & Jim McMullan.
393255	135501	Green eggs and ham, by Dr. Seuss [pseudonym]
379848	2452011	There is a bird on your head! / by Mo Willems.

```
SELECT COUNT(*) AS TotalCount,
YEAR(from_unixtime(unix_timestamp(CheckoutDateTime,"MM/dd/yyyy hh:mm:ss aaa")))
AS year
FROM lab04data
GROUP BY YEAR(from_unixtime(unix_timestamp(CheckoutDateTime,"MM/dd/yyyy hh:mm:ss aaa")))
ORDER BY year;
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter columns

totalcount	year
49135	null
3793697	2005
6580531	2006
7089632	2007
8371667	2008
9032881	2009
8417704	2010
7773942	2011
7292070	2012
7859162	2013
7413761	2014
6869813	2015
6403183	2016
5033515	2017