Siraj Bagwan

Project on Big Data Platform Engineering

Challenges

- Improve click-through rates

Improving click-through rate (the number of clicks that your content receives divided by the number of times your content is shown.) by recommending content to relevant users.

- Personalize online ads and content

Showing more relevant ads and content basis on their past interaction to online users to improve user experience.

- Lead acquisition

By improving user experience and interaction can be able to raise the ratio of subscription.

Solution

- Collect

collect and gather data of user behavior and interaction and store it to data lake from different sources in order to process.

- Process

Transform the click stream data on data lake and ingest it to data warehouse. Ingest the data from OLTP system to Data Warehouse.

- Analyze

Analyze the data present on the cluster using different tools available in order to generate insights.

Capacity Planning

Capacity Planning		
Daily Data in Motion:	90 GB	90 GB
Replication Factor:	3	
Total Data in a Day (Daily data in motion * replication) :	90 * 3=270 GB	270 GB
Data in a Month:	270 *30 = 8100 GB	7.91 TB
Data in Year:	7.91*12 = 115.92 TB	94.92 TB
Data at Rest:	30 TB	30 TB
Replication of data at rest	30*3=90	90 TB
DFS Data:		**184.92 TB**
10% Overhead Needed:	10% of (94.92)	**9.43 TB**
Non DFS Data and Overhead 30% :	30% of (184.92)	**55.47 TB**
Final Data:	184.92+9.43+55.47=249.82	**~250 TB**

- Data retention period: 1 year
- Total data collected per year: 250 TB
- Data to be stored on each data node: 10 TB
- No. of data nodes required: 250/10 = 25
- 10% overhead (node failure): 3
- Total data nodes required: 28
- Kafka nodes: 3

Cluster Planning

Hosts	No. of Hosts Required	Specification
Master Hosts	3	Instance Type: r6a.4xlarge Ram: 128 GB Core: 16
Utility Hosts	2	Instance Type: r6a.4xlarge Ram: 128 GB Core: 16
Edge hosts	1	Instance Type: c6a.8xlarge Ram: 64 GB Core: 32
Worker Hosts	28	Instance Type: c5a.16xlarge Ram: 128 GB Core: 64

Cluster Planning

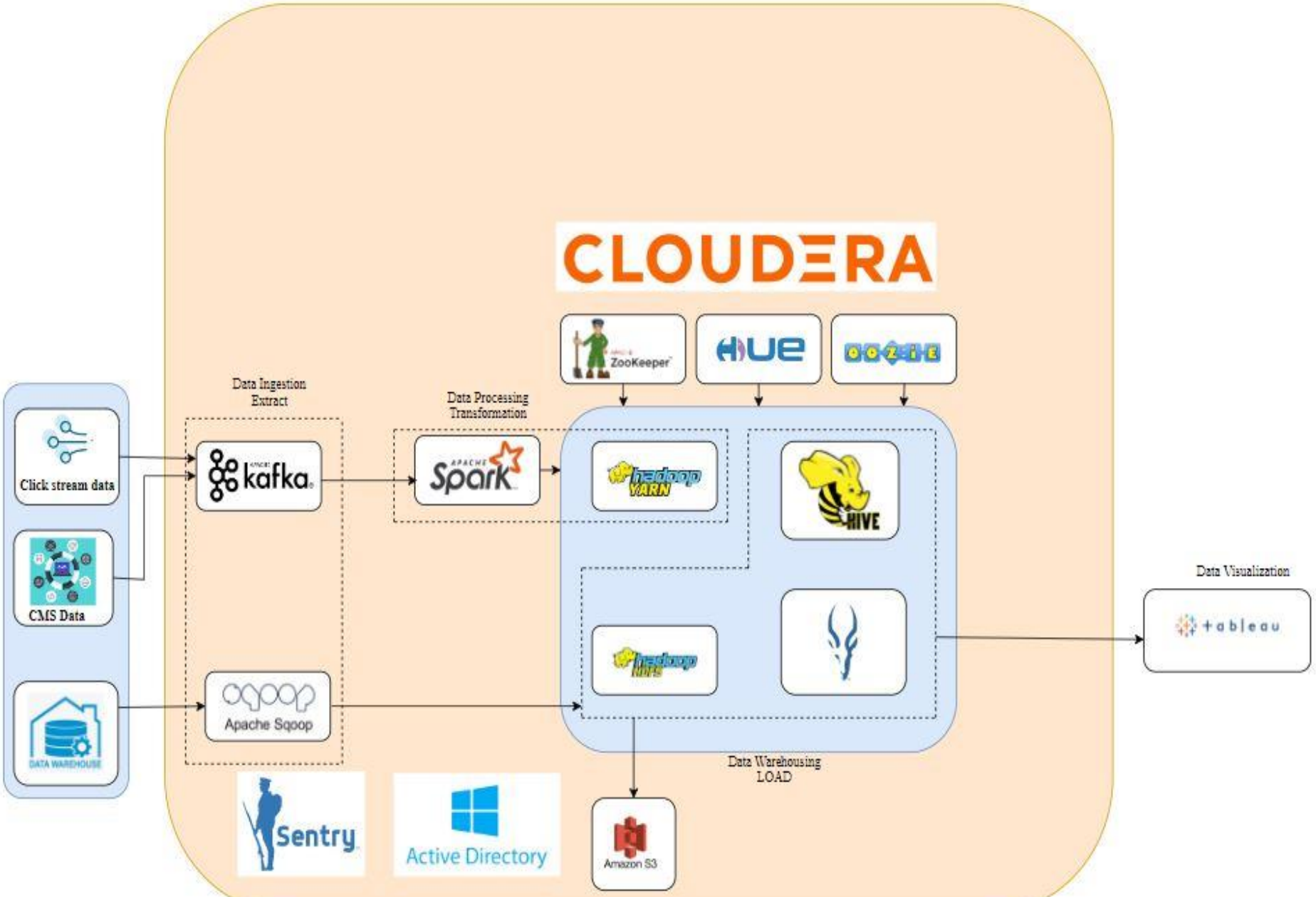
Nodes	Services
Master Node 1:	NN, JN, Failover Controller, Zookeeper Resource Manager,
Master Node 2:	Standby NN, JN, Failover Controller, Standby Resource Manager, Zookeeper
Master Node 3:	JN, Zookeeper, JHS,SHS
Utility Node 1:	Cloudera Manager
Utility Node 2:	HMS,HS2,ICS,SS,
Edge Node:	Gateway of HDFS, YARN, HIVE. HUE, OOZIE
Data Nodes:	DN, NM, ID
3 Kafka Nodes:	Kafka Brokers

Cluster Planning

- Block size of HDFS: 128 MB
- 1 MB fsimage size for per 1000 blocks (Cloudera Suggest)
- 250 TB = 256000 GB = 262,144,000 MB
- No. of blocks = 262144000/128 =2048000
- Fsimage = 2048000/1000 = 2048 MB = 2GB
- Heap Size of name node = 2 GB * 2 = 4GB

Service Stack

Services	Versions
Hadoop	3.0.0
Kafka	2.1.0
Sqoop	1.4.7
Spark	2.4.0
Hive	2.1.1
Impala	3.2.0
Hue	4.3.0
Oozie	5.1.0
Zookeeper	3.4.5
Sentry	2.1.0



Data Flow Diagram

Cloudera Manager | Clusters | Hosts | Diagnostics | Audits | Charts | Backup | Administration | Search | Support | admin

Home

Status | All Health Issues | Configuration | All Recent Commands | Switch to Table View | Add

Production

CDH 6.2.0 (Parcels)

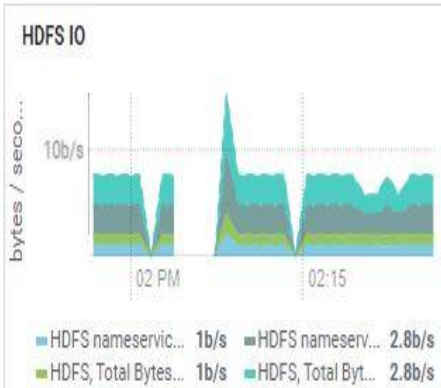
- 14 Hosts
- HDFS
- Hive
- Hue
- Impala
- Kafka
- Oozie
- S3 Connector
- Sentry
- Spark
- Sqoop 1 Client
- YARN (MR2 In...
- ZooKeeper

Charts

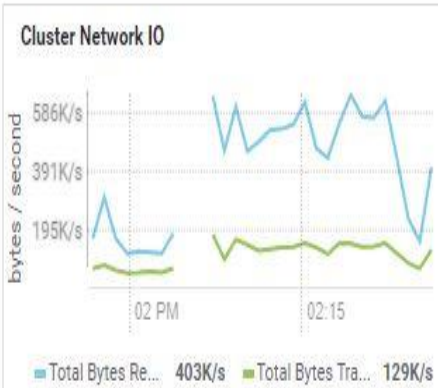
30m 1h 2h 6h 12h 1d 7d 30d

Completed Impala Queries

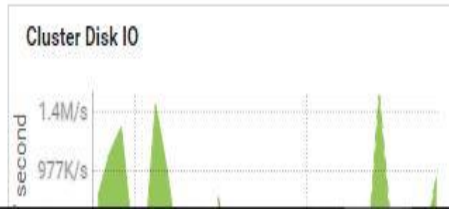
HDFS IO



Cluster Network IO



Cluster Disk IO



Cluster CPU

Activate Windows
Go to Settings to activate Windows.

Feedback

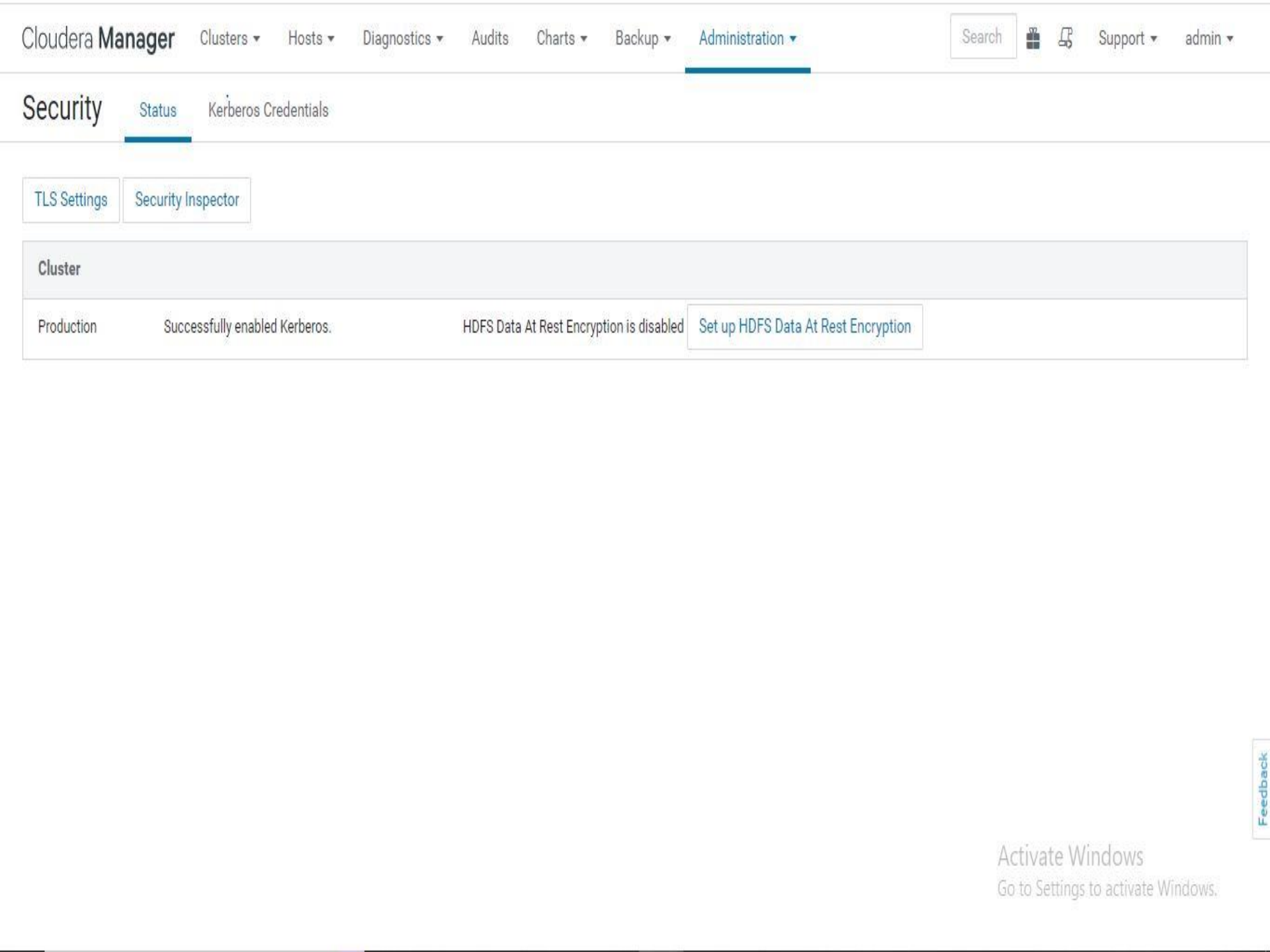
Roles

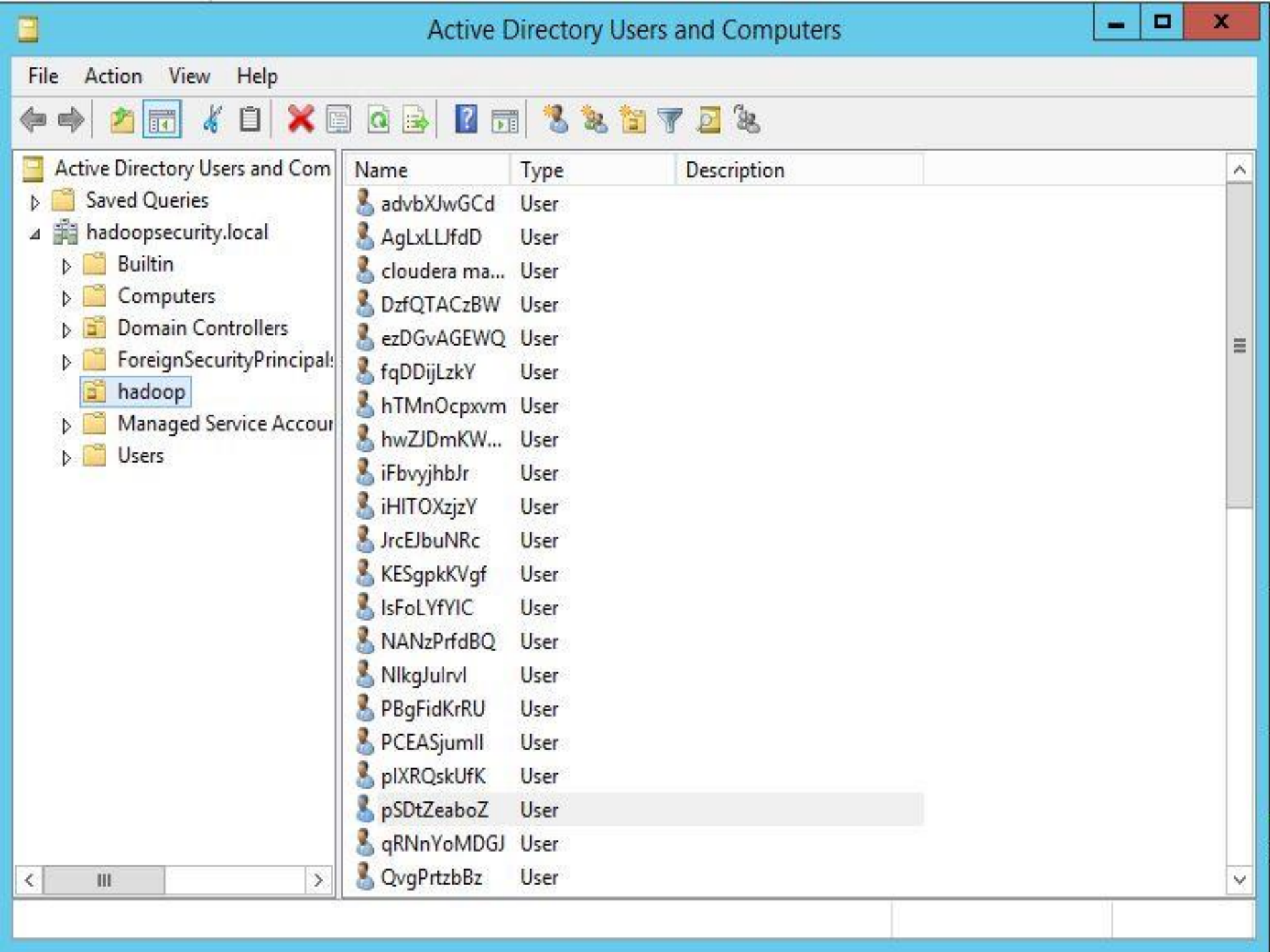
Hosts	Count	Roles
ip-10-0-0-105.ap-south-1.compute.internal	1	FC JN NN RM S
ip-10-0-0-113.ap-south-1.compute.internal	1	G G LB HS KTR G OS G G G G
ip-10-0-0-120.ap-south-1.compute.internal	1	JN G SS HS JHS S
ip-10-0-0-121.ap-south-1.compute.internal	1	B FC JN NN RM S
ip-10-0-0-122.ap-south-1.compute.internal	1	AP ES HM RM SM
ip-10-0-0-89.ap-south-1.compute.internal	1	G HMS HS2 ICS ISS G
ip-10-0-0-[69, 82, 84, 104, 115].ap-south-1.compute.internal	5	DN ID NM
ip-10-0-0-[75, 77, 88].ap-south-1.compute.internal	3	KB

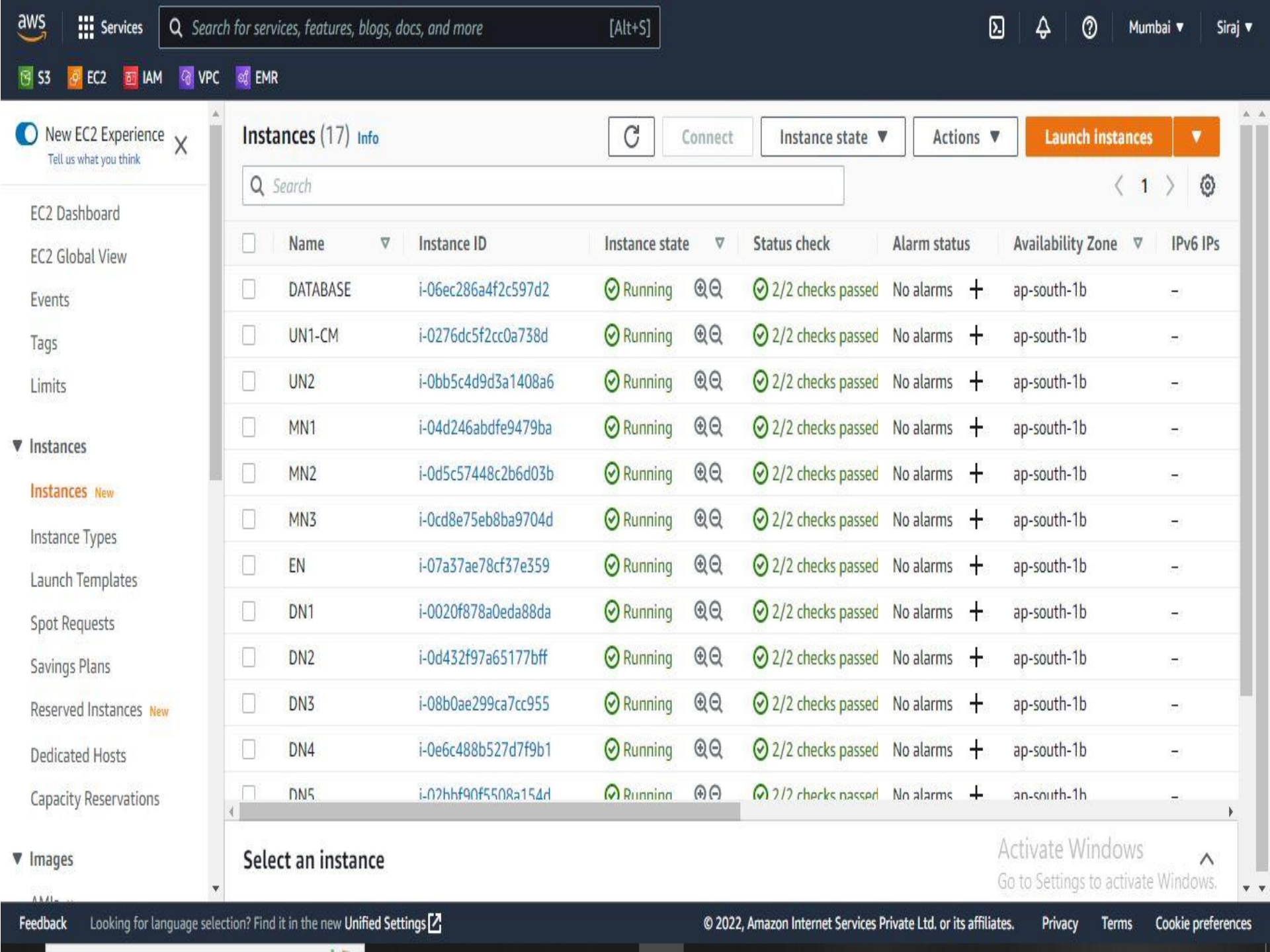
This table is grouped by hosts having the same roles assigned to them.

Feedback

Activate Windows
Go to Settings to activate Windows.


Cloudera Manager
Clusters
Hosts
Diagnostics
Audits
Charts
Backup
Administration
Search
Support
admin
Security
Status
Kerberos Credentials
TLS Settings
Security Inspector
Cluster
Production
Successfully enabled Kerberos.
HDFS Data At Rest Encryption is disabled
Set up HDFS Data At Rest Encryption
Feedback
Activate Windows
Go to Settings to activate Windows.


Active Directory Users and Computers
File
Action
View
Help
Active Directory Users and Com
Saved Queries
hadoopsecurity.local
Builtin
Computers
Domain Controllers
ForeignSecurityPrincipal
hadoop
Managed Service Accour
Users
Name
Type
Description
advbXJwGCd User
AgLxLLJfdD User
cloudera ma... User
DzfQTACzBW User
ezDGvAGEWQ User
fqDDijLzkY User
hTMnOcpxvm User
hwZJDmKW... User
iFbvyjhbJr User
iHITOXzjzY User
JrcEJbuNRc User
KESgpkKVgf User
IsFoLYfYIC User
NANzPrfdBQ User
NIkgJulrvl User
PBgFidKrRU User
PCEASjumll User
plXRQskUfK User
pSDtZeaboZ User
qRNnYoMDGJ User
QvgPrtzbBz User


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai
Siraj
S3
EC2
IAM
VPC
EMR
New EC2 Experience
Tell us what you think
EC2 Dashboard
EC2 Global View
Events
Tags
Limits
Instances
Instances New
Instance Types
Launch Templates
Spot Requests
Savings Plans
Reserved Instances New
Dedicated Hosts
Capacity Reservations
Images
Instances (17) Info
Connect
Instance state
Actions
Launch instances
Search
1
Name
Instance ID
Instance state
Status check
Alarm status
Availability Zone
IPv6 IPs
DATABASE i-06ec286a4f2c597d2 Running 2/2 checks passed No alarms ap-south-1b –
UN1-CM i-0276dc5f2cc0a738d Running 2/2 checks passed No alarms ap-south-1b –
UN2 i-0bb5c4d9d3a1408a6 Running 2/2 checks passed No alarms ap-south-1b –
MN1 i-04d246abdfe9479ba Running 2/2 checks passed No alarms ap-south-1b –
MN2 i-0d5c57448c2b6d03b Running 2/2 checks passed No alarms ap-south-1b –
MN3 i-0cd8e75eb8ba9704d Running 2/2 checks passed No alarms ap-south-1b –
EN i-07a37ae78cf37e359 Running 2/2 checks passed No alarms ap-south-1b –
DN1 i-0020f878a0eda88da Running 2/2 checks passed No alarms ap-south-1b –
DN2 i-0d432f97a65177bff Running 2/2 checks passed No alarms ap-south-1b –
DN3 i-08b0ae299ca7cc955 Running 2/2 checks passed No alarms ap-south-1b –
DN4 i-0e6c488b527d7f9b1 Running 2/2 checks passed No alarms ap-south-1b –
DN5
Select an instance
Activate Windows
Go to Settings to activate Windows.
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

Impact

- Click through rate increased thoroughly
- Five times increase in conversion rate for successful subscriptions
- Grew profitability with delivery of new services and customized content
- Increased subscription sales

THANK YOU