

Big Data Platform Assessment

Steps to run an HDI Assessment & an HDP Questionnaire

Prepared by

Data SQL Ninja Engineering Team (datasqlninja@microsoft.com)

This document is provided "as-is". Information and views expressed in this document, including URL and other Internet Web site references, may change without notice.

Some examples depicted herein are provided for illustration only and are fictitious. No real association or connection is intended or should be inferred.

This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

© 2019 Microsoft. All rights reserved.

Note: The detail provided in this document has been harvested as part of a customer engagement sponsored through the [Data SQL Ninja Engineering](#).

Table of Contents

Introduction	4
1 Assessment objectives and scope	5
1.1 Objectives.....	5
1.1.1 General Assessment scope.....	5
2 Assessment approach	6
2.1 Assess Head Node Size.....	6
2.2 Assess Core usage	7
2.3 Assess primary storage	8
2.4 Assess metastore settings	9
2.5 Assess Hive usage and settings.....	9
2.6 Assess Oozie usage and settings.....	11
2.7 Review Persisted Script Actions.....	11
3 HDP to HDI Migration Questionnaire	12
3.1 Project Details.....	12
3.2 Environment.....	13
3.3 Security.....	14
3.4 Security Preferences.....	14
3.5 Source Data	15
3.6 Metadata	15
3.7 Scale	16
3.8 Cluster Utilization.....	16
3.9 Current Limitations.....	16
3.10 Data Movement.....	17
3.11 Monitoring & Alerting	17
3.12 Architecture Preferences	17
3.13 Azure Infrastructure	18
3.14 Staff.....	18
4 Feedback and suggestions.....	19

Introduction

This whitepaper is a cheat sheet on how to quickly assess a customer's HDInsight environment and assess an on-premise HDP solution. The steps outline the process to collect data on the cluster, its configuration and its usage. This allows the Architect to efficiently and thoroughly perform a detailed migration to Cloud and an HDInsight Assessment exercise.

1 Assessment objectives and scope

1.1 Objectives

The core objectives of this Assessment are:

- Review current HDI implementation landscape.
- Review current HDI Cluster configuration.

1.1.1 General Assessment scope

The scope of the assessment will be to:

- Review current HDI Cluster performance
 - Review data flow
 - Review batch sizes
 - Review read & write speeds
- Review current HDI Cluster configuration:
 - Primary storage
 - Metastore settings
 - Hive, Oozie, and Ranger DB settings
 - Worker node size
 - Head node size
 - Core usage limits
 - Persisted script actions

2 Assessment approach

The assessment consists of several steps. The steps can all be taken from the Azure Portal with proper access rights to the HDInsight Cluster. The admin username & password are needed in order to be able to access the Ambari views. The steps of the assessment are:

1. Assess Head & Worker Node sizes
2. Assess Core usage
3. Assess primary storage
4. Assess metastore settings
5. Assess Hive usage and settings
6. Assess Oozie usage and settings
7. Review persisted script actions

2.1 Assess Head Node Size

The first step is to assess the head and worker node sizes configured for the cluster. This can be seen in the "Overview" pane below. You can see the cluster has 6 nodes with 2 head and 4 worker nodes.

The screenshot shows the Azure Portal interface for an HDInsight cluster named 'hdinsighttestingip'. The left sidebar contains a navigation menu with options like Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Tools, Settings, Cluster size, Quota limits, SSH + Cluster login, Data Lake Storage Gen1, Storage accounts, Applications, Script actions, External metastores, HDInsight partner, Properties, Locks, and Export template. The main content area displays cluster details: Resource group (change) : deletemeasp, Status : Running, Location : East US 2, Subscription (change) : Visual Studio Enterprise, Subscription ID : 404dfd6e-f9d8-41a7-8524-b5a45640e85e, and Tags (change) : Click here to add tags. Below this, there's a 'Cluster dashboards' section with links to Ambari home, Ambari views, Zeppelin notebook, Jupyter notebook, Spark history server, and Yarn. The 'Cluster size' section shows a table with 6 nodes:

TYPE	SIZE	CORES	NODES
Head	D12 v2	8	2
Worker	D13 v2	32	4

Going to the “Cluster Size” tab under “Settings” you can see the exact cost of the current configuration.

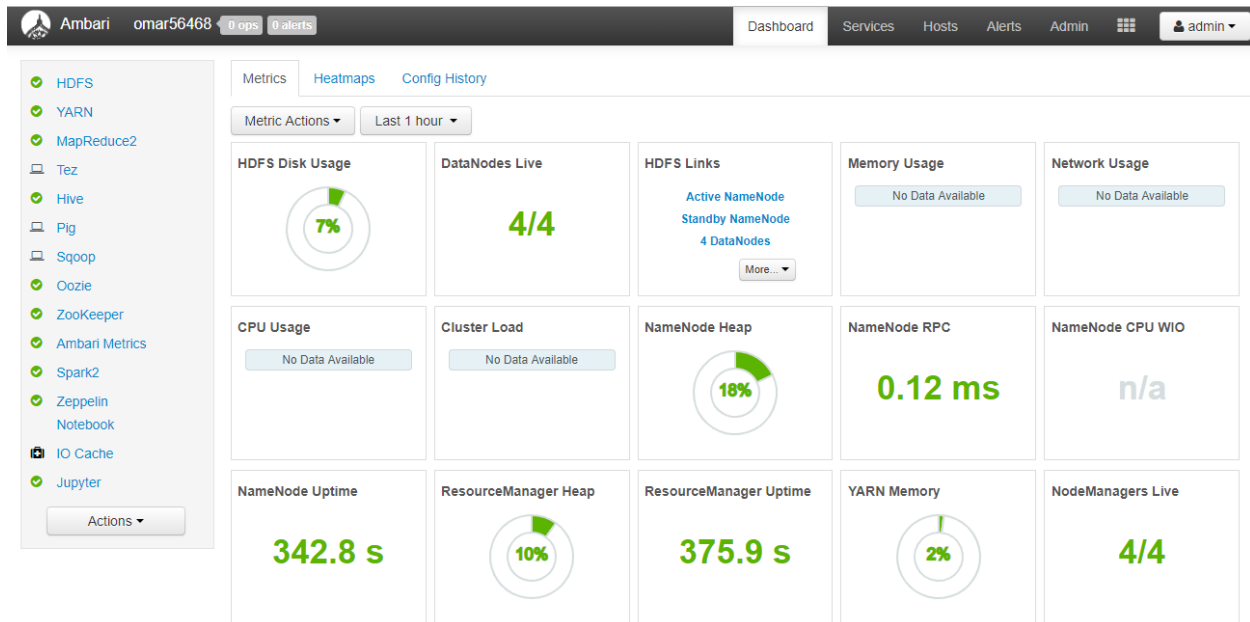
The screenshot shows the 'Cluster size' configuration page for an HDInsight cluster named 'omar56468'. The left sidebar contains navigation links: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Tools, Settings, Cluster size (selected), Quota limits, SSH + Cluster login, Data Lake Storage Gen1, and Storage accounts. The main content area shows the 'Number of Worker nodes' set to 4. Below this, it lists 'Worker node sizes' as D13 v2 (4 nodes, 32 cores) and 'Head node size' as D12 v2 (2 nodes, 8 cores). A table summarizes the costs: WORKER NODES at 0.748 x 4 = 2.990, HEAD NODES at 0.374 x 2 = 0.748, and a TOTAL COST of 3.74 USD/HOUR (ESTIMATED). A note states: 'This configuration will use 40 of 20 available cores in the East US 2 region. Need more cores? Contact support.' A disclaimer at the bottom reads: 'This estimate does not include subscription discounts or costs related to storage, networking, or data transfer.'

2.2 Assess Core usage

To see the core usage, we navigate to the “Ambari home” under the “Cluster dashboards” control in the “Overview” page.

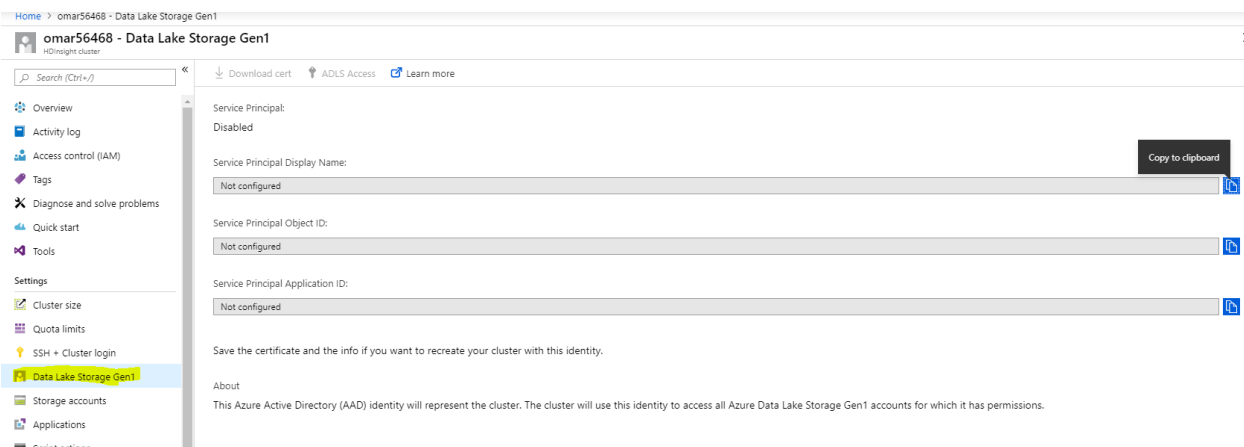
The screenshot shows the 'Overview' page for the HDInsight cluster 'omar56468'. The left sidebar is the same as in the previous screenshot. The main content area displays cluster details: Resource group (change) : delete, Status : Running, Location : East US 2, Subscription (change) : Microsoft Azure Internal Consumption, Subscription ID : 0516e634-0741-41cf-a437-1420fe25661f, and Tags (change) : Click here to add tags. On the right, there are links for 'Learn more : Documentation', 'Cluster type, HDI version : Spark 2.3 (HDI 3.6)', 'URL : https://omar56468.azurehdinsight.net', and 'Getting started : Quickstart'. A 'Cluster dashboards' section is highlighted, listing 'Ambari home', 'Ambari views', 'Zeppelin notebook', 'Jupyter notebook', 'Spark history server', and 'Yarn'.

Here you can see the HDFS Disk Usage, DataNodes Live, Memory Usage, Network Usage, CPU Usage along with many other metrics.

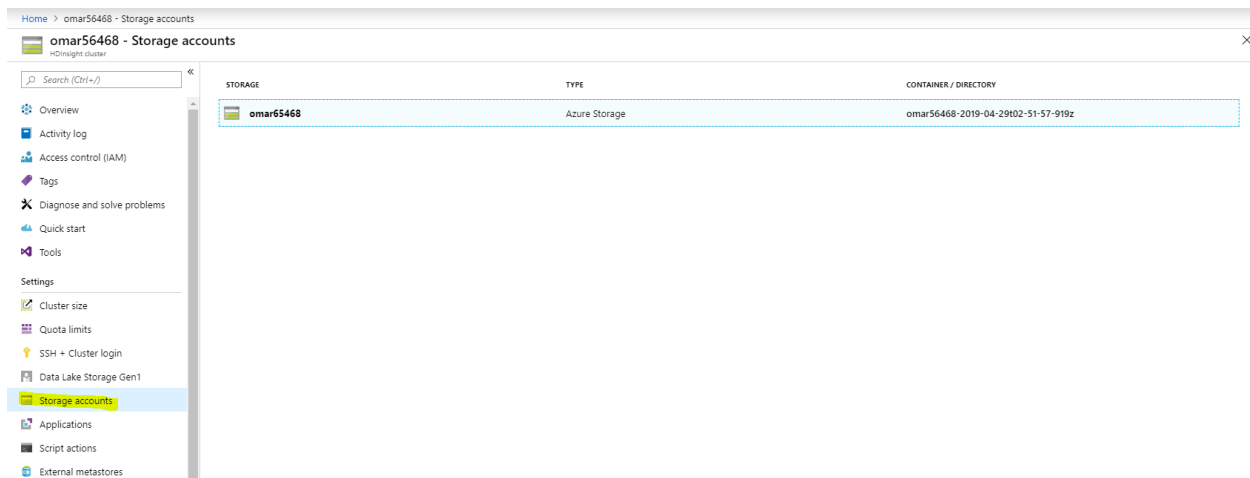


2.3 Assess primary storage

For HDInsight, ADLS or SA can be used as primary storage. This can easily be assessed by going to the Data Lake Storage Gen1 tab to see if it is configured or “Not Configured”.

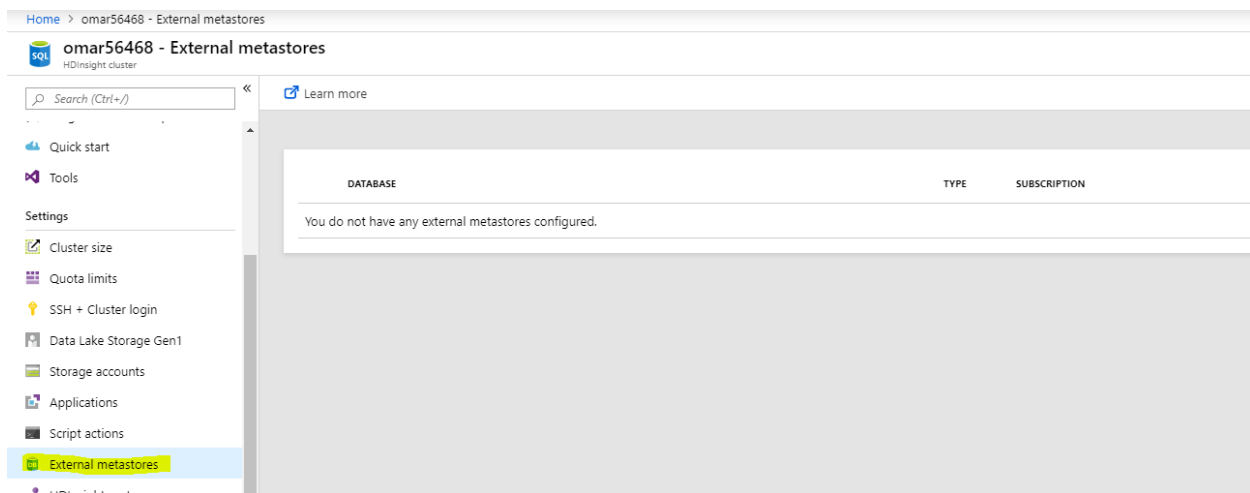


The next step is to go to “Storage Accounts” to see the container configured.



2.4 Assess metastore settings

Assessing External Metastores is as straight-forward as the Primary storage. To make sure the cluster isn't using any external metastores, we just need to navigate to the "External Metastores" tab. The External Metastores are where metadata can be stored in Azure SQL DB.



2.5 Assess Hive usage and settings

From the Ambari views, we can navigate to Hive to see the summary and configuration of our Hive repository. From here, we can also drill into the metastores and servers to gather more information if needed.

Ambari omar56468 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin

Summary Configs Service Actions

Summary No alerts

- [Hive Metastore](#) Started No alerts
- [Hive Metastore](#) Started No alerts
- [HiveServer2](#) Started No alerts
- [HiveServer2](#) Started No alerts
- [WebHCat Server](#) Started No alerts
- [WebHCat Server](#) Started No alerts
- [HCat Client](#) 1 HCat Client Installed
- [Hive Clients](#) 6 Hive Clients Installed

HiveServer2 JDBC URL jdbc:hive2://zk1-omar56.lfira0nknuffclukeurxyt5g.cx.internal.cloudapp.net:2181,zk...

Hive View 2.0 [Go To View](#)

Debug Hive Query [Go To View](#)

Actions

Also by navigating to the Yarn dashboard, we can use the query and application running and whether there are failed jobs or not.

hadoop All Applications Logged in as: dr:who

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
3	0	2	1	2	3 GB	200 GB	0 B	2	60	0	4	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[MEMORY]	<memory 512, vCores 1>	<memory 51200, vCores 15>

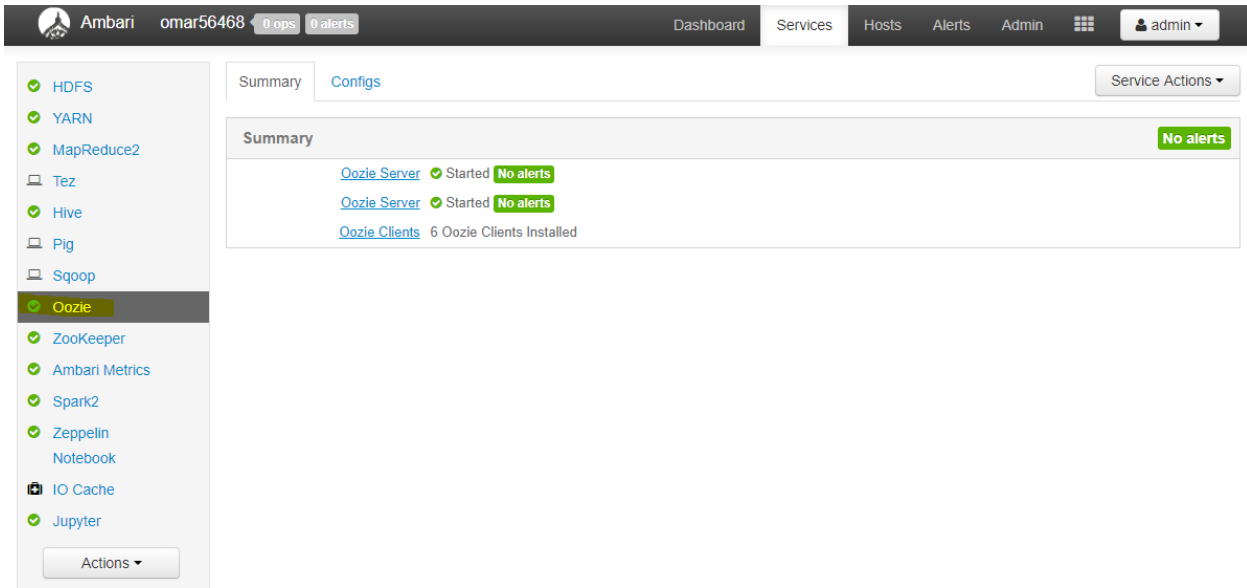
Show 20 entries

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
application_1556506914531_0003	root	HIVE-37921b4-f5a6-4a36-b73a-554d15dddbbe	TEZ	default	0	Mon Apr 29 13:03:38 +1000 2019	Mon Apr 29 13:03:58 +1000 2019	FINISHED	SUCCEEDED	N/A	N/A	N/A	0.0	0.0		History	N/A
application_1556506914531_0002	hive	Thrft JDBC/ODBC Server	SPARK	thrifsvr	0	Mon Apr 29 13:03:09 +1000 2019	N/A	RUNNING	UNDEFINED	1	1	1536	1.5	0.8		ApplicationMaster	0
application_1556506914531_0001	hive	Thrft JDBC/ODBC Server	SPARK	thrifsvr	0	Mon Apr 29 13:02:58 +1000 2019	N/A	RUNNING	UNDEFINED	1	1	1536	1.5	0.8		ApplicationMaster	0

Showing 1 to 3 of 3 entries

2.6 Assess Oozie usage and settings

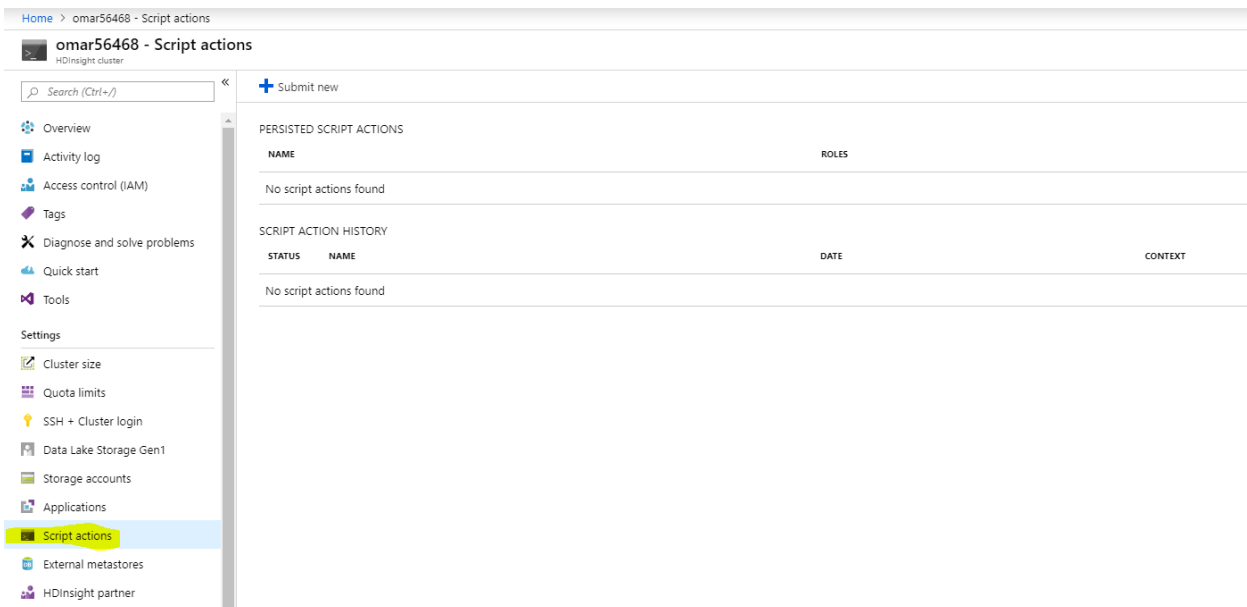
As with Hive, navigating to Oozie allows us to see the Oozie usage and settings.



The screenshot shows the Ambari web interface. The top navigation bar includes the Ambari logo, the cluster name 'omar56468', and status indicators for '0 ops' and '0 alerts'. The main navigation menu on the left lists various services: HDFS, YARN, MapReduce2, Tez, Hive, Pig, Sqoop, Oozie (highlighted), ZooKeeper, Ambari Metrics, Spark2, Zeppelin Notebook, IO Cache, and Jupyter. The 'Services' tab is selected in the top bar. The Oozie service summary shows 'Oozie Server' as 'Started' with 'No alerts' and 'Oozie Clients' as '6 Oozie Clients Installed'.

2.7 Review Persisted Script Actions

The last major step of the assessment is to review the Persisted Script Actions. Persisted Scripts are scripts that could be performing business logic using the cluster. This is essential for understanding the workloads that will be migrated. The scripts can be seen from the “Script actions” tab under “Settings”.



The screenshot shows the 'Script actions' page in the Ambari Settings section. The left sidebar lists various settings categories, with 'Script actions' highlighted. The main content area shows a search bar and a 'Submit new' button. Below this, there are two sections: 'PERSISTED SCRIPT ACTIONS' and 'SCRIPT ACTION HISTORY'. Both sections currently display 'No script actions found'.

3 HDP to HDI Migration Questionnaire

The following questionnaire is to be filled out prior to any migration from HDP to HDI. It covers the main aspects required in order to properly size the HDI environment and architect the solution.

3.1 Project Details

The purpose of these questions is to collect data on what the project will require in terms of features and architectural pieces.

Question	Example	Answer
MapReduce Jobs	10 Jobs (Twice Daily)	
Hive jobs	100 jobs (Every Hour)	
Spark batch jobs	50 jobs (Every 15 Minutes)	
Spark Streaming jobs	5 jobs -- every 3 minutes	
Structured Streaming jobs	5 jobs -- every minute	
ML Model training jobs	2 jobs -- once in a week	
Programming Languages	Python, Scala, Java	
Scripting	Shell, Python	

3.2 Environment

This section is for understanding the current environment so a parallel environment can be architected.

Question	Example	Answer
Cluster Distribution type	Hortonworks, Cloudera, MapR	
Cluster Distribution version	HDP 2.6.5, CDH 5.7	
Big Data eco-system components	HDFS, Yarn, Hive, LLAP, Impala, Kudu, HBase, Spark, MapReduce, Kafka, Zookeeper, Solr, Sqoop, Oozie, Ranger, Atlas, Falcon, Zeppelin, R	
Cluster types	Hadoop, Spark, Confluent Kafka, Storm, Solr	
Number of clusters	4	
Number of Master Nodes	2	
Number of Worker Nodes	100	
Number of Edge Nodes	5	
Total Disk space	100 TB	
Master Node configuration	m/y, cpu, disk, etc.	
Data Nodes configuration	m/y, cpu, disk, etc.	
Edge Nodes configuration	m/y, cpu, disk, etc.	
HDFS Encryption?	Yes	
High Availability	HDFS HA, Metastore HA	
Disaster Recovery / Back up	Backup cluster?	
Systems that are dependent on Cluster	SQL Server, Teradata, Power BI, MongoDB	
Third-party integrations	Tableau, GridGain, Qubole, Informatica, Splunk	

3.3 Security

The existing security provisions used on HDP are covered here in order to understand the required security on Azure.

Question	Example	Answer
Perimeter security	Firewalls	
Cluster authentication & authorization	Active Directory, Ambari, Cloudera Manager, No authentication	
HDFS Access Control	Manual, ssh users	
Hive authentication & authorization	Sentry, LDAP, AD with Kerberos, Ranger	
Auditing	Ambari, Cloudera Navigator, Ranger	
Monitoring	Graphite, collectd, statsd, Telegraf, InfluxDB	
Alerting	Kapacitor, Prometheus, Datadog	
Data Retention duration	3 years, 5 years	
Cluster Administrators	Single Administrator, Multiple Administrators	

3.4 Security Preferences

The required security infrastructure is queried here. This is for the "to-be" environment on Azure.

Question	Example	Answer
Private and protected data pipeline?	Yes	
Domain Joined cluster (ESP)?	Yes	
On-Premises AD Sync to Cloud?	Yes	
No. of AD users to sync?	100	

Ok to sync passwords to cloud?	Yes
Cloud only Users?	Yes
MFA needed?	No
Data authorization requirements?	Yes
Role-Based Access Control?	Yes
Auditing needed?	Yes
Data encryption at rest?	Yes
Data encryption in transit?	Yes

3.5 Source Data

Source data information is collected here in order to understand the existing data and how it is transported.

Question	Example	Answer
Data sources	Flat files, Json, Kafka, RDBMS	
Data orchestration	Oozie workflows, Airflow	
In memory lookups	Apache Ignite, Redis	
Data destinations	HDFS, RDBMS, Kafka, MPP	

3.6 Metadata

The metadata store is covered here in order to understand how metadata is currently handled.

Question	Example	Answer
Hive DB type	MySQL, PostgreSQL	
No. of Hive Metastores	2	
No. of Hive tables	100	
No. of Ranger policies	20	

No. of Oozie workflows	100
------------------------	-----

3.7 Scale

The sizing for the data, clusters and ingestion volumes is covered here.

Question	Example	Answer
Data volume including Replication	100 TB	
Daily ingestion volume	50 GB	
Data growth rate	10% per year	
Cluster Nodes growth rate	5% per year	

3.8 Cluster Utilization

Cluster usage and utilization is covered here in order to understand the future cluster size that may be needed in order to save costs and ensure maximum performance.

Question	Example	Answer
Average CPU % used	60%	
Average Memory % used	75%	
Disk space used	75%	
Average Network % used	25%	

3.9 Current Limitations

Any limitations/drivers for migration are covered here. List problems with the current environment.

Question	Example	Answer
Current limitations	Latency is high	
Current challenges	Concurrency issue	

3.10 Data Movement

This section covers how data is currently transported in HDP.

Question	Example	Answer
Initial load preference	DistCp, Data box, ADF, WANDisco	
Data transfer delta	DistCp, AzCopy	
Ongoing incremental data transfer	DistCp, Sqoop	

3.11 Monitoring & Alerting

Please list your preferences for Monitoring & Alerting here.

Question	Example	Answer
Use Azure Monitoring & Alerting Vs Integrate third-party monitoring	Use Azure Monitoring & Alerting	

3.12 Architecture Preferences

If there are Architecture preferences, please fill them out below. Microsoft will provide the architectural best practices.

Question	Example	Answer
Single cluster vs Specific cluster types	Specific cluster types	
Colocated Storage Vs Remote Storage?	Remote Storage	
Smaller cluster size as data is stored remotely?	Smaller cluster size	
Use multiple smaller clusters rather than a single large cluster?	Use multiple smaller clusters	
Use a remote metastore?	Yes	

Share metastores between different clusters?	Yes
Deconstruct workloads?	Replace Hive jobs with Spark jobs
Use ADF for data orchestration?	No

3.13 Azure Infrastructure

The preferred requirements for the new Azure environment.

Question	Example	Answer
Preferred Region	US East	
VNet preferred?	Yes	
HA / DR Needed?	Yes	
Integration with other cloud services?	ADF, CosmosDB	

3.14 Staff

The current staff members dedicated to this migration.

Question	Example	Answer
No. of Administrators	2	
No. of Developers	10	
No. of end users	100	
Skills	Hadoop, Spark	
No. of available resources for Migration efforts	2	

4 Feedback and suggestions

If you have feedback or suggestions for improving this data migration asset, please contact the Data SQL Ninja Team (datasqlninja@microsoft.com). Thanks for your support!

Note: For additional information about migrating various source databases to Azure, see the [Azure Database Migration Guide](#).