DARE
Installation Guide

In order for DARE to operate, a two-step installation is required. The two steps include the installation of CDH and Apache Spot. CDH is the most complete distribution of Apache Hadoop and related projects. Apache Spot is an *open-source software for leveraging insights from flow and packet analysis*.

CDH (Cloudera Manager / Hadoop / Spark)

3 Ubuntu 16.04 VMs running on ESXi hosts.

- First VM is the Cloudera Manager host and runs Hadoop master processes such as HDFS, YARN, etc.
- Second VM hosts Cloudera Edge Node and essentially acts as the client access point for launching jobs in the cluster.
- Third VM hosts Cloudera Worker and runs DataNodes and other distributed processes.

Hosts	# of CPUs	Memory (GB)	Disk (GB)	Roles
Cloudera Manager	8	20	400	· HDFS DataNode · HDFS HttpFS · HDFS NFS Gateway · HDFS SecondaryNameNode · Hive Gateway · Hive MetaStore Server · HiveServer2 · Impala Daemon · Kafka Gateway · Kafka MirroMaker · Activity Monitor · Alert Publisher · Event Server · Host Monitor · Service Monitor · Spark Gateway · YARN JobHistory Server · YARN NodeManager · YARN ResourceManager

Cloudera Edge Node	8	20	400	· HDFS Balancer · HDFS NFS Gateway · HDFS NameNode · Hive Gateway · Impala Catalog Server · Impala StateStore · Kafka Gateway · Kafka Broker · Spark Gateway · ZooKeeper Server
Cloudera Worker	8	20	400	· HDFS DataNode · HDFS NFS Gateway · Hive Gateway · Impala Daemon · Kafka Gateway · Kafka Broker · Spark Gateway · Spark History Server · YARN NodeManager

CDH environment setup and configuration

The installation procedure is sufficiently documented in Cloudera's installation guide and requires 7 main steps as follows:

Step 1: Configure a Repository:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/configure_cm_repo.html#cm_repo

Step 2: Install JDK:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/cdh_ig_jdk_installation.html#topic_29

Step 3: Install Cloudera Manager Server:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/install_cm_cdh.html#cmig_topic_6_6

Step 4: Install Databases:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/cm_ig_installing_configuring_dbs.html#cmig_topic_5

Step 5: Set up the Cloudera Manager Database:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/prepare_cm_database.html#cmig_topic_5_2

Step 6: Install CDH and Other Software:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/install_software_cm_wizard.html#cm_installation_wizard

Step 7: Set Up a Cluster:
https://www.cloudera.com/documentation/enterprise/6/6.0/topics/cluster_setup_wizard.html#concept_b4d_wkh_ycb

Apache Spot

To install the latest stable version of Apache Spot, clone the following link (as tarball):

https://www-us.apache.org/dist/incubator/spot/1.0-incubating/apache-spot-1.0-incubating.tar.gz.

Alternatively, our forked repository[1] which contains all relevant DARE modules, including Apache Spot can be cloned.

You can verify this release using 1.0 signatures and checksums [PGP, SHA-512, MD5] with project release KEYS.

In order to validate the build, follow the instructions:

- Download the tarball from the above link.
- Decompress the tarball:
  tar -zxvf apache-spot-1.0-incubating.tar.gz
- Change directory:
  cd apache-spot-1.0-incubator.

Since Apache Spot is composed of more than one module or sub-projects, some of them need compilation, while others (Python or Javascript based) don't.

To install each module please follow the below instructions:

- Decompressed tarball content should be the same with the content located in:
  https://github.com/apache/incubator-spot/tree/v1.0-incubating
- To install each component please follow the official documentation guide:
  http://spot.apache.org/doc/#installation

[1] https://github.com/shield-h2020/dare

Spot-ingest (centralized), Spot-setup, Spot UI and Spot-OA have specific requirements to install manually.
http://spot.apache.org/doc/#configuration
http://spot.apache.org/doc/#ingest
http://spot.apache.org/doc/#oa
http://spot.apache.org/doc/#ui

Spot-ML is the only component to build the binary files using sbt assembly commands. To install it follow these instructions:
http://spot.apache.org/doc/#ml

Components are also available https://github.com/shield-h2020.

Distributed collector – Streaming Worker

Distributed collector is a daemon that runs in each vNSF. Its installation includes dependencies in Python regarding:

- Avro serialization framework.
- Kafka – Python for the Apache Kafka distributed system.
- Watchdog – a Python API and shell utilities that monitor file system events.

Also, in Linux OS, there is a dependency with pip, the Python package manager, and in order to process specific pipelines installation of appropriate tools is needed:

- Spot - nfdump – a version for processing netflows.
- Tshark – a part of wireshark distribution for processing pcap files.

All the necessary installation and dependency files, along with detailed configuration can be found at SHIELD repositories.[2][3]

Helpful comments

In order for everything to operate as smoothly as possible, it is highly recommended to refresh package handler across all three nodes. In our Ubuntu installation these commands were used:

- sudo apt-get update

[2] https://github.com/shield-h202/vnsfs-collectors

[3] https://github.com/shield-h2020/vnsfs-workers

- sudo apt-get upgrade.

Furthermore, in order to make access in all three nodes more convenient we have created a new user called “spotuser”, which also is a sudoer, thus making installation easier as well.