LIGADATA

Fatafat 1.0 Setup and Running Guide

Overview

This guide will help you get the LigaDATA Fatafat engine (Online Learning Engine Platform) and Metadata API installed and ready to use.

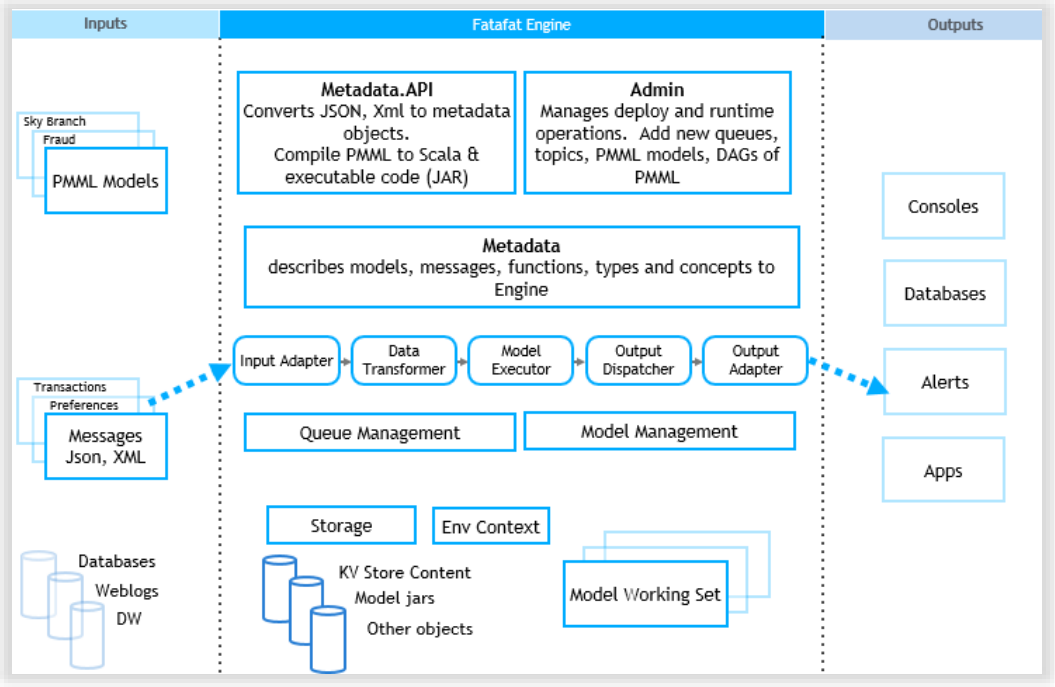
High Level Process Steps in this Guide:

Step 1. Installing Fatafat
1. Install Fatafat Package from Source OR
2. Install Fatafat Package from Binaries

Step 2. Deploying and Running Fatafat
1. Load Core Fatafat Metadata
2. Load Application Metadata
3. Load the input data and the Rule set
4. Create Queues and Push data to Kafka Queues
5. Run the engine
6. Push Sample Data
7. View Results

Once you have started the Engine, it will process the input data against the rules defined in the Metadata and produce "alerts" (output that matches your application's rules).

The Fatafat engine processes incoming streaming data, transforms it into messages or containers, and processes it according to the ruleset you have supplied in your models. It then produces decision or alert data that can be acted upon. Below, you will find a high level diagram of the components.



MetadataAPI Objects

The metadata objects include Types, Functions, Concepts, Message Definitions, and Model Definitions. The Metadata API defines create and read operations for metadata objects supported by this system. All functions take String values as input in XML or JSON Format and return a JSON string of ApiResult object.

Prerequisites:

- CentOS/RedHat/OS X (If using windows, set up a virtual machine running a Linux distribution)
- Approximately 400 MB for installation (3 GB if building from source)
    - Access to: https://github.com/ligaDATA/Fatafat OR the Fatafat install package
- Install the JDK 1.7.1 or greater (which can be downloaded here.)
- Install Scala v2.10.4: http://www.scala-lang.org/download/2.10.4.html
- Install sbt: http://www.scala-sbt.org/download.html
- Download and install Zookeeper: http://zookeeper.apache.org/releases.html#download
- Download and install Kafka 2.10-0.8.1.1: http://kafka.apache.org/
- The definition of the data models and rule sets used in your business that need to be created as models.

Assumptions:

At least one instance of each of the following is running:

- Zookeeper Service
- Kafka
- Cassandra OR HBase

NOTE: Cassandra and HBase are optional - you will need some input/output to the engine.

Linux users: It is suggested that Linux users run the services/scripts under screens (http://www.howtoforge.com/linux_screen), so that if the connection is lost, they will still be running. This will likely become a service in a later release.)

Step 1: Installing Fatafat

Note: Install from source (option 1.) or binary (option 2.)

Option 1: Install Fatafat Package from Source:

1. Download the project source.
    a. From GitHub, clone: https://github.com/ligaDATA/Fatafat.git
2. Run install script:
    a. cd <localCloneDirectory>/Fatafat/trunk/SampleApplication/EasyInstall
    b. bash easyInstallFatafat.sh <InstallDirectoryPath> <trunk sources Path> <ivyPath for dependency jars> <KafkaLocation>

Example: bash easyInstallFatafat.sh /tmp/InstallDirectory ~/repos/Fatafat/trunk ~/.ivy2 ~/Kafka/kafka_2.10-0.8.1.1/kafka_2.10-0.8.1.1

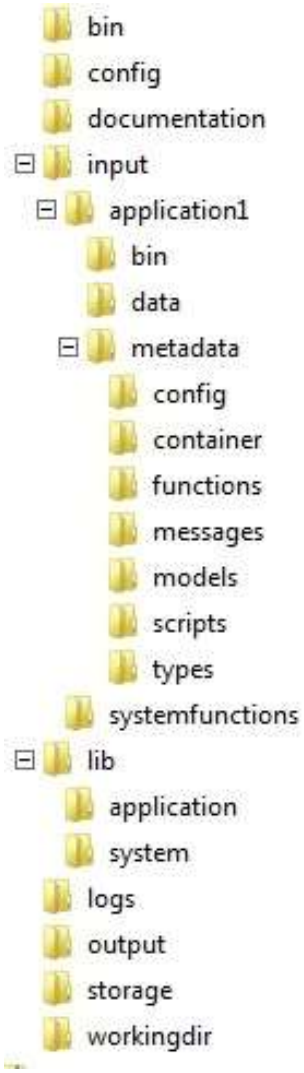
Option 2: Install Fatafat Package from Binaries:

1. For this example, we will assume the binaries are on a thumb drive. Copy the Fatafat directory from the thumb drive to a local directory (Referred to hereafter as <InstallDirectoryPath>).

2. Run the command below to set the needed paths to generate the scripts from the current directory:

    a. bash <InstallDirectoryPath>/bin/SetPaths.sh

After installation, the following directory structure will be created:



Step 2: Deploying and Running Fatafat

Note: You will need to ensure that Zookeeper and Kafka are running before running Fatafat.

#1 - Load Core Fatafat Metadata

1. Execute the following script to load cluster configuration into metadata:
    a. bash <InstallDirectoryPath>/bin/ClusterMetadata.sh
    b. Add Cluster Configuration (Option 33)

#2 - Load Application Metadata

1. Execute the following script to load the application definitions into metadata:
    a. bash <InstallDirectoryPath>/input/application1/bin/ApplicationMetadata.sh
    b. Add Containers, Messages, Models

    NOTE: Each application will have an individual script for loading metadata

#3 - Load the input data and the Rule set

1. Execute the following command to load the patient data (containers) and the rule set:
    a. bash <InstallDirectoryPath>/input/application1/bin/InitKvStores.sh

#4 - Create Queues and Push data to Kafka Queues

1. Execute the following command to create queues:
    a. bash <InstallDirectoryPath>/bin/CreateQueues.sh

#5 - Run the engine:

1. Execute the following command to run the Engine:
    a. bash <InstallDirectoryPath>/bin/StartEngine.sh
    b. Typing “quit” or pressing ctrl-c will stop the engine

#6 - Push Sample Data:

1. Execute the following command to push input data to queues:
    a. bash <InstallDirectoryPath>/input/application1/bin/PushSampleDatazToKafka.sh
    b. The input file being used can be found at
    <InstallDirectoryPath>/input/application1/data/copd_demo.csv.gz

#7 - View Results

1. Execute the following script to see status:
    a. bash <InstallDirectoryPath>/bin/WatchStatusQueue.sh
2. Execute the following script to see the output results:
    a. bash <InstallDirectoryPath>/bin/WatchOutputQueue.sh

THE REMAINING CONTENT STILL NEEDS REVIEW.

Multi-Node Account Setup Instructions

Each Fatafat cluster installed on a network is ideally run by a distinct user account. These instructions provide the basics for setting up such a user so that it can easily manage the cluster administration.

A user name and group should be selected to be the owner and user that runs the cluster. It can be any useful and meaningful name not already in use. In the examples below, the name *fatafat* is used for both the account and group name. Two nodes named *ls19.dc.npario.com* and *ls20.dc.npario.com* are used to illustrate the setup on two machines.

Installation:

#1 – Administration Set Up Actions

Verify or do each of the following.

If these are already done, or you are working on a previous installation, these should be confirmed. This assumes that you can get to these machines from another account and have sudoer privileges.

Create the Fatafat Admin group

1. If the *fatafat* group does not exist on the prospective cluster nodes, add it to each node with the following command:

   sudo groupadd fatafat

Create a User Account on each Cluster

2. Establish a user account on each prospective cluster node making the new user use the fatafat group as its default login group. Be sure to establish a home directory for the user. Example:

   sudo useradd --gid fatafat --create-home fatafat

3. Set the password for the fatafat cluster login to something using sudo

   sudo passwd fatafat

Ensure the staging user/developer is a member of the *fatafat* group

4. If the fatafat release is staged from source on one of the machines in the new cluster or cluster's network, make sure that the staging user (or the developer) account is also a member of the *fatafat* group. This can be done with this command:

   sudo usermod -G fatafat releaseAdmin

Here the user account is releaseAdmin. This really is only needed on this one machine, not every node in the cluster, although that would not necessarily be a bad idea. This is security policy dependent – ensure that you have those policies correctly setup.

Ensure passphrase-less access to all nodes

5. Passphrase-less access to/from all participating nodes in the *fatafat* cluster is required for proper operation. Each node could have its own public and private key if your security policy requires it. However, it is far easier to login to one of the nodes as the fatafat user and set up the .ssh keys once, then propagate the entire directory to the other nodes in the cluster. Assuming the accounts are new, there is really no harm in replacing the entire .ssh directory. If you have existing keys there, however, don't replace. Instead append.

Establish Public/Private key

1. generate a public private key

   ssh-keygen

It will prompt for a key type. Take the default (RSA 2048) by hitting return. Then it will ask for a passphrase; hit return for "passphraseless" behavior. It will prompt again to type the same passphrase; once again hit return. This will generate a public and private key for the current user account.

2. Install the public key on the current machine.

To install the new public key in the current directory, you can cat the public key into a file called .ssh/authorized_keys or perhaps a little easier use this approach, as it will properly set the access permissions for the file it will create. If the current machine is named fatafat19.distCluster.com and the fatafat account being setup is fatafat, run the following command:

   ssh-copy-id fatafat@fatafat19.distCluster.com

You will have to type the password this one time to establish the authorized key file, but after that you should be able to access without giving a passphrase.

3. Verify that the $HOME/.ssh permissions is strict. There should be no permitted access to the .ssh directory except by the owner.

Create the .ssh archive to set up remaining nodes with the correct .ssh keys

4. While operating as the fatafat admin account, Zip/Bzip/Gzip the .ssh folder in its entirety and leave it in $HOME. For example:

tar cvjf fatafatNodeSSH.bz2 .ssh

This creates an archive that will be copied around to the other nodes that are to participate in the cluster.

5. As the fatafat account, copy the archive to every node in the cluster and unzip it. For example, if fatafat20.distCluster.com were another node in the cluster to be setup, do the following:

   scp fatafatNodeSSH.bz2 fatafat20.distCluster.com:

This will leave the tarball in the fatafat user's fatafat20 home folder. Do this same action for all the nodes that participate or might participate in the cluster being set up.

6. Tar the ssh to each node as the fatafat user and decompress the file copied in the previous step. For example,

   tar xvjf fatafatNodeSSH.bz2

This will update the .ssh folder with the authorized_keys as well as the public and private keys that were generated in part a) above. To verify that it worked, try to ssh to the original machine where the keys were generated. For example,

   ssh fatafat20.distCluster.com

If successful, then you should login immediately (not requiring a passphrase from the console).

7. Check further. A good check would be to login to one of the cluster nodes that will participate in the fatafat cluster being formed as the fatafat account and try to log in to every other machine in the cluster. These logins should succeed without a passphrase being required.

Should there be problems, there are many resources on the web that describe the setup of a passphrase-less account. In particular, the access permissions for the account and its .ssh need to be strict (no group or global permissions should be allowed). Ssh will complain if these requirements are not met.

8. Finally, remove the tarball that has been passed around to all of the nodes from each fatafat user account on each node.

More information. The instructions above are indicative, not authoritative, of what must be done to establish password-less access. If you experience problems, consult the documentation for your Linux enlistment for special requirements that your environment may require.

Installing Fatafat to a cluster from your Git Repository

Download and Invoke the Scripts that set up the MetadataAPI configuration, the KVStores and the sample Kafka data

Fatafat is typically installed from the trunk directory of an "install" or "release" machine, building the Fatafat installation directory on the local system, creating a tarball of that directory, sending it to the nodes that are part of the Fatafat cluster, untar and decompress the tarball there and move the resulting directory into place.

Several scripts are part of the repository that are used to accomplish this. These scripts must be moved onto your PATH and given "execute" permissions. They are:

1. trunk/SampleApplication/Medical/template/config/MetadataAPIConfig_Template.properties
2. trunk/SampleApplication/Medical/template/script/InitKvStores_Template.sh
3. trunk/SampleApplication/Medical/template/script/PushSampleDataToKafka_Template.sh
4. trunk/Pmml/Scripts/sbtProjDependencies/src/main/scala/sbtProjDependencies.scala

The ***clusterInstallFatafat.sh*** *script* controls the installation, using the cluster configuration information either found in the supplied NodeConfigPath argument or from the metadata store described in the MetadataAPIConfig file. In the first case, the metadata store may have no configuration and the supplied NodeConfigPath will provide it. In the second case, where no NodeConfigPath is supplied, the metadata store is expected to have one. Should a valid configuration not be found, the script will issue a RuntimeException with a useful message describing the particular issue. Here are examples of script invocation:

clusterInstallFatafat.sh
--MetadataAPIConfig SampleApplication/Medical/Configs/MetadataAPIConfig.properties

clusterInstallFatafat.sh
--NodeConfigPath SampleApplication/Medical/Configs/Engine2BoxConfigV1.json

The ***installFatafat_Medical.sh*** will build the Fatafat install directory. The *nodeInfoExtract.*sh is used to determine which cluster nodes are to receive the build. The s*btProjDependencies* is a useful tool used by a number of scripts that are part of Fatafat. In this case, it is used to build the classpath for the *nodeInfoExtract* `application.

Invoke the installer

Use the following command by invoking it from the trunk directory of the local Git repository which is to be installed. Run it from the account that is to be used to run/manage the Fatafat cluster and refer to directories that are write-able by that account.

```
Cluster InstallFatafat.sh
  --MetadataAPIConfig  SampleApplication/Medical/Configs/MetadataAPIConfig.properties
  --NodeConfigPath SampleApplication/Medical/Configs/Engine2BoxConfigV1.json
```

Currently two configuration files are required, one for the MetadataAPI and the other that describes at least the cluster for the FatafatManager. The cluster description is the key information as to what sort of distribution and what machines and nodes are involved in the installation. Note that multiple nodes can if desired be installed on the same physical computer.

For more information about what these configuration files contain, see <some reference to other documentation describing the configuration>.

Start the Cluster

The Fatafat cluster start script - **startFatafatCluster.sh**, like the install script, can take one or two arguments. If the NodeConfigPath is present, its cluster config will be added to the metadata store specified with the MetadataAPIConfig argument. If not present, the metadata store is expected to have a configuration. Should a valid configuration not be found, the script will issue a RuntimeException with a useful message describing the particular issue.

```
startFatafatCluster.sh
--MetadataAPIConfig  <metadataAPICfgPath> [--NodeConfigPath <fatafatCfgPath> ]
```