**Hadoop Deployment Layout**

## With increased complexity and evolving Hadoop ecosystem, having standard deployment layout ensures better integration between Hadoop sub-projects. By making the installation process easier, we can lower the barrier to entry and increase Hadoop adoption.

## Packages

We need to divide Hadoop up into packages that can be independently upgraded. The list of packages should include:

* Hadoop Common - Common including the native code and required jar files.
* HDFS Client - HDFS jars, scripts, and shared libraries.
* HDFS Server - jsvc executable
* Yarn Client - Yarn client jars and scripts
* Yarn Server - Yarn server jars and scripts
* MapReduce - MapReduce jars, scripts, and shared libraries
* LZO - LZ0 codec from github.com/omally/hadoop-gpl-compression
* Metrics - Plugins for Chukwa and Ganglia

Packages from other teams will include:

* Pig
* Hive
* Oozie client
* Oozie server
* Howl client
* Howl server

## Deployment

It is important to have a standard deployment that results from installing the packages regardless of the package manager. Here are the top level directories and a sample of what would be under each. Note that all of the packages are installed "flattened" into the prefix directory. For compatibility reasons, we should create "share/hadoop" that matches the old HADOOP\_HOME and set the HADOOP\_HOME variable to that.

$PREFIX/ bin / hadoop

| | mapred

| | pig -> pig7

| | pig6

| + pig7

|

+ etc / hadoop / core-site.xml

| | hdfs-site.xml

| + mapred-site.xml

|

+ include / hadoop / Pipes.hh

| | + TemplateFactory.hh

| + hdfs.h

|

+ lib / jni / hadoop-common / libhadoop.so.0.20.0

| |

| | libhdfs.so -> libhdfs.so.0.20.0

| + libhdfs.so.0.20.0

|

+ libexec / task-controller

|

+ man / man1 / hadoop.1

| | mapred.1

| | pig6.1

| + pig7.1

|

+ share / hadoop-common

| | hadoop-hdfs

| | hadoop-mapreduce

| | pig6

| + pig7

|

+ sbin / hdfs-admin

| | mapred-admin

|

+ src / hadoop-common

| | hadoop-hdfs

| + hadoop-mapreduce

|

+ var / lib / data-node

| + task-tracker

|

| log / hadoop-datanode

| + hadoop-tasktracker

|

+ run / hadoop-datanode.pid

+ hadoop-tasktracker.pid

## Path Configurations

Path can be configured at compile phase or installation phase. For RPM, it takes advantage of the --relocate directive to allow path reconfiguration at install phase. For Debian package, path is configured at compile phase.

Build phase parameter:

* package.prefix - Location of package prefix (Default /usr)
* package.conf.dir - Location of configuration directory (Default /etc/hadoop)
* package.log.dir - Location of log directory (Default /var/log/hadoop)
* package.pid.dir - Location of pid directory (Default /var/run/hadoop)