Summary

The current efforts to improve YARN scheduler leads to the requirement for propagating additional information from AM to RM. For example, YARN-2745 requires the AM to propagate information such as estimated task duration, total number of tasks, selectivity factor, sensitivity to locality, disk and network rates.

Having a description of a group of tasks, it allows the RM scheduler to take better decisions based on the various characterizations of the groups.

In this JIRA, we propose enhancing the AM RM protocol with a new type of message. Every group of tasks will be tagged with a group label accordingly.

Design

We propose to enhance the AllocateRequest datastructure with a list of profiles. More precisely we propose to modify:

yarn\_protos.proto to include an additional message:

message GroupProfileProto {

optional string group\_label = 1;

optional int32 num\_tasks = 2;

optional int32 duration = 3;

optional float selectivity = 4;

optional float nw\_read\_rate = 5;

optional float nw\_write\_rate = 6;

optional float disk\_read\_rate = 7;

optional float disk\_write\_rate = 8;

optional bool locality\_sensitive = 9 [default = false];

}

and AllocateREquestProto in yarn\_service\_protos.proto to include a list of GroupProfileProto:

message AllocateRequestProto {

repeated ResourceRequestProto ask = 1;

repeated ContainerIdProto release = 2;

optional ResourceBlacklistRequestProto blacklist\_request = 3;

optional int32 response\_id = 4;

optional float progress = 5;

repeated ContainerResourceIncreaseRequestProto increase\_request = 6;

**repeated GroupProfileProto group\_profiles = 7;**

}.

In addition, we enhance the ResourceRequestProto with an additional group\_label field:

message ResourceRequestProto {

optional PriorityProto priority = 1;

optional string resource\_name = 2;

optional ResourceProto capability = 3;

optional int32 num\_containers = 4;

optional bool relax\_locality = 5 [default = true];

optional string node\_label\_expression = 6;

**optional string group\_label = 7;**

}

such that every ask will be tagged with a corresponding group label profile which enables the RM to use additional information at the time of scheduling.