What kinds of configuration exists in Hive on Spark?

In Hive driver side:
1. Hive configuration. e.g. hive.execution.engine.
2. HDFS configuration. e.g. dfs.namenode.https-address.
3. YARN configuration. e.g. yarn.resourcemanager.address.
4. Spark configuration. e.g. spark.master.
5. RSC configuration. e.g. hive.spark.client.rpc.max.size.

In Spark side (include RemoteDriver):
1. YARN configuration.
2. Spark configuration.
3. RSC configuration.

I separate RSC configuration from Hive configuration as it required both in Hive driver side and RemoteDriver side. Spark cluster does not depends on Hive configuration to launch, although the spark job of Hive queries does, and Hive configuration is serialized as part of SparkJob.

How to set configurations in Hive on Spark?

For Hive driver configuration (priority from high to low):
1. SET command.
2. Hive CLI option.
3. Configuration file.

Currently we support load spark configuration from hive configuration file and spark configuration file both, and spark configuration file priority is higher than hive configuration.

For Spark configuration (priority from high to low):
1. RemoteDriver option. (by Hive)
2. Spark submit option. (by Hive)
3. Configuration file.

What configuration should be pushed from Hive to Spark?

We should only need to push the configuration which is related to spark cluster launch, other configuration like Hive configuration and HDFS configuration is required in Spark at job level, and them would be transferred to Spark cluster together with job.
For RSC mode, Spark need:

1. YARN configuration.
2. Spark configuration.
3. RSC configuration.

For local Spark context mode, Spark only need:

1. YARN configuration.
2. Spark configuration.

How to push configuration from Hive to Spark?

1. Spark configuration. Hive could add Spark configuration through SparkSubmit option, we use –properties-file now.
2. RSC configuration. RemoteDriver support –conf option, RSC configuration could be pushed through this option.
3. YARN configuration. Append “spark.hadoop.” prefix as YARN configuration property name, and add to SparkConf same as #1, Spark would remove the prefix and take it as Hadoop configuration.

NOTE: Due to the implementation of Spark, several configurations may only enabled in certain deploy mode, HIVE-9342 is an example.

How to make dynamic configuration setting work for Hive on Spark?

As we know, Spark cluster is launched in application level instead of job level. If we want to update Spark configuration, we need to re-launch a new Spark cluster, which means that, in Hive on Spark, we should create new SparkSession to support dynamic configuration setting while user set spark/YARN/RSC configuration.

What do we still miss by now?

1. Clean up work, move all RSC, Spark configuration (in Hive) to HiveConf for consistent management.

2. Push RSC configuration to Spark side in RSC mode.
3. Push YARN configuration to Spark side in Spark on YARN mode.
4. Reset SparkSession while RSC/YARN configuration updated (we already done that for Spark configuration update).