SK C&C Big Data Knowledge Test:

<Part I>

0. Challenges

    a. Create and place your work in challenges folder in your GitHub project.
    b. All screenshot must be in PNG format
    c. All text file (including command set) require Markdown(.md) formatting.
    d. You will create each required file yourself
    e. Make sure you provide screenshots and/or cut and paste of your command line outputs for each of the items that you are asked to do.
    f. It is ok to combine multiple requirements into a single screenshot and/or cut/paste of your terminal

1. Create a CDH Cluster on AWS

You will be given 5 public and private IP addresses along with a public access key to 5 instances running on AWS.

Connect to your instances and install a Cloudera CDH cluster.

    a. Linux setup
        i. Add the following linux accounts to all nodes
            1. User training with a UID of 3800
            2. Set the password for user “training” to “training”
            3. Create the group skcc and add training to it
            4. Give training sudo capabilities
        ii. List the your instances by IP address and DNS name (don’t use /etc/hosts for this)
        iii. List the Linux release you are using
        iv. List the file system capacity for the first node (master node)
        v. List the command and output for yum repolist enabled
        vi. List the /etc/passwd entries for training (only in master name node)
        vii. List the /etc/group entries for skcc (only in master name node)
        viii. List output of the flowing commands:
            1. getent group skcc
            2. getent passwd training

    b. Install a MySQl server
        i. Use MariaDB as the database for all the services. You may choose your own username and passwords but make a record of it so that we may access them.
        ii. List the following in your GitHub

1. A command and output that shows the hostname of your database server
2. A command and output that reports the database server version
3. A command and output that lists all the databases in the server

c. Install Cloudera Manager
   i. Specifically, you MUST install CDH version 5.15.2 You will lose points if you install any other version of CDH.
   ii. The Cluster does not have to be in HA mode.
   iii. Make sure that the following services (and any necessary services to install that service) are installed:
      1. HDFS
      2. YARN
      3. Sqoop
      4. Hive
      5. Impala
      6. HUE
   iv. In you cluster, create a user named "training" with password "training"
      1. You should have already created the linux user from previous step. Now, make sure user "training" has both a linux and HDFS home directory

2. In MySQL create the sample tables that will be used for the rest of the test
   a. In MySQL, create a database and name it "test"
   b. Create 2 tables in the test databases: authors and posts.
      i. You will use the authors.sql and posts.sql script files that will be provided for you to generate the necessary tables
   c. Create and grant user "training" with password "training" full access to the test database. (It is ok if you give training full access to the entire MySQL database)

3. Extract tables authors and posts from the database and create Hive tables.
   a. Use Sqoop to import the data from authors and posts
   b. For both tables, you will import the data in tab delimited text format
   c. The imported data should be saved in training's HDFS home directory
      i. Create authors and posts directories in your HDFS home directory
      ii. Save the imported data in each
   d. In Hive, create 2 tables: authors and posts. They will contain the data that you imported from Sqoop in above step.
   e. You are free to use whatever database in Hive.
   f. Create authors as an external table.
   g. Create posts as a managed table.

4. Create and run a Hive/Impala query. From the query, generate the results dataset that you will use in the next step to export in MySQL.
   a. Create a query that counts the number of posts each author has created.
      i. The id column in authors matches the author_id key in posts.
   b. The output of the query should provide the following information:

Source	Output Column Name
Id from authors	Id
first_name from authors	fname
last_name from authors	Lname
Aggregated count of number of posts	num_posts

   c. The output of the query should be saved in your HDFS home directory.
      i. Save it under “results” directory

5. Export the data from above query to MySQL
   a. Create a MySQL table and name it “results”
      i. Make sure it has the necessary columns of matching type as the results of your query from above
   b. The table should be created under the database “test”
   c. Finally, export into MySQL the results of your query