<https://github.com/kpasva/HadoopSpark.git>

**TASK 01 - SPARK SUBMIT**

Write the Spark command to submit spark application on YARN and mode is client. Class name: SparkYarn, Jar location: /home/horton/solutions/SparkYarn.jar

*Save the script to your GIT link in the path /YOUR\_NAME/task01*

**TASK 02 - Convert an existing HDFS file to ORC format and save it in HDFS**

* Data file is available in GIT ‘SOURCE DATA’ folder file system under TASK02
* Fields (stockticker:string, transactiondate:string, openprice:float, highprice:float, lowprice:float, closeprice:float, volume:bigint)
* Convert file format to ORC
* Save the script to your GIT link in a text file in the path */YOUR\_NAME/task02*
* Save the output file to the path- /user/<YOUR\_NAME>/nyse\_orc

**TASK 03- Convert an existing HDFS file to Parquet format and save it in HDFS**

* Data file is available in GIT ‘SOURCE DATA’ folder file system under TASK02 (Use the same file as used in TASK 02)
* Fields (stockticker:string, transactiondate:string, openprice:float, highprice:float, lowprice:float, closeprice:float, volume:bigint)
* Convert file format to ORC
* Save the script to your GIT link in a text file in the path */YOUR\_NAME/task03*
* Save the output file to the path- /user/<YOUR\_NAME>/nyse\_parquet

**TASK 04 - Calculate and find the numbers of words ‘ERROR’ in the file using Spark Accumulator and display in the terminal**

* Data file is available in GIT ‘SOURCE DATA’ folder file system under TASK04
* Using SPARK ACCUMULATORS calculate the number of time the word ‘ERROR’ is available in the file
* Display the count in the Terminal
* Save the script to your GIT link in a text file in the path */YOUR\_NAME/task04*

**TASK 05 - Get top 3 crime types based on number of incidents in RESIDENCE area**

* Data file is available in GIT ‘SOURCE DATA’ folder file system under TASK05
* Structure of data (ID,Case Number,Date,Block,IUCR,Primary Type,Description,Location Description,Arrest,Domestic,Beat,District,Ward,Community Area,FBI Code,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location) File format - text file
* Delimiter - “,” (use regex while splitting split(",(?=(?:[^\"]\*\"[^\"]\*\")\*[^\"]\*$)", -1), as there are some fields with comma and enclosed using double quotes.
* Get top 3 crime types based on number of incidents in RESIDENCE area using “Location Description”
* Store the output result in HDFS path /user/<YOUR\_NAME>/solutions/solution03/RESIDENCE\_AREA\_CRIMINAL\_TYPE\_DATA
* Save the script to your GIT link in a text file in the path */YOUR\_NAME/task05*
* Output Fields: Crime Type, Number of Incidents
* Output File Format: TXT
* Output Delimiter: N/A

**TASK 06 - Load a JSON file in Hive and using Spark write a program for the below given scenario**

* Data is available in GIT ‘SOURCE DATA’ folder file system under Exercise 06
* Load the JSON file into Hive
* Using Spark Data frames calculate the average count of babies by year across the counties sort by gender
* Save the script to your GIT link in a text file in the path */YOUR\_NAME/task06*
* Store the result as tab separated records in a new directory in HDFS named /user/<YOUR\_NAME>/task06