

CA675 Assignment2 For Spark

Step 1: Startup a cluster in GCP

The screenshot shows the Google Cloud Platform console interface. The top navigation bar includes the Google Cloud Platform logo, a search bar, and a user profile icon. The left sidebar contains a menu with options like 'Jobs on clusters', 'Clusters', 'Jobs', 'Workflows', 'Autoscaling policies', 'Utilities', 'Component exchange', 'Metastore', and 'Release notes'. The main content area displays the 'Cluster details' for a cluster named 'cluster-cd23'. The cluster is in a 'Running' status. Below the details, there are tabs for 'MONITORING', 'JOBS', 'VM INSTANCES', 'CONFIGURATION', and 'WEB INTERFACES'. The 'VM INSTANCES' tab is selected, showing a table of instances:

Name	Role
cluster-cd23-m	Master
cluster-cd23-w-0	Worker
cluster-cd23-w-1	Worker

Below the table, there is a link for 'EQUIVALENT REST'.

Step 2: Upload CSV file to HDFS

The screenshot shows a terminal window with the following commands and output:

```
zhe_wang23@cluster-cd23-m: ~ - Google Chrome
ssh.cloud.google.com/projects/cloud-technology-assignment/zones/europe-west2-a/instances/cluster-cd23-m?authuse
zhe_wang23@cluster-cd23-m:~$ hdfs dfs -mkdir /user/crimedata
zhe_wang23@cluster-cd23-m:~$ hdfs dfs -cp gs://hive-test1/cirme/crime.csv /user/crimedata
zhe_wang23@cluster-cd23-m:~$ hdfs dfs -ls /user/crimedata
Found 1 items
-rw-r--r-- 2 zhe_wang23 hadoop 2302 2021-11-16 22:55 /user/crimedata/crime.csv
zhe_wang23@cluster-cd23-m:~$
```

Step 3: Launch spark-shell

```
zhe_wang23@cluster-cd23-m: ~ - Google Chrome
```

```
ssh.cloud.google.com/projects/cloud-technology-assignment/zones/europe-west2-a/instances/cluster-cd23-m?authse
```

```
zhe_wang23@cluster-cd23-m:~$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://cluster-cd23-m.c.cloud-technology-assignment.internal:41291
Spark context available as 'sc' (master = yarn, app id = application_1637100259326_0002).
Spark session available as 'spark'.
Welcome to

      ____  

     / ___/  

    /  _<__\    

   /  / |___ \    

  /  /___)  >   

 /_____/    

version 2.4.8

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_275)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Step 4: Load CSV File into Spark and check filed name

```
scala> import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.SQLContext

scala> val sqlContext = new SQLContext(sc)
warning: there was one deprecation warning (since 2.0.0): for details, enable `:setting -deprecation` or `:replay -deprecation`
sqlContext: org.apache.spark.sql.SQLContext = org.apache.spark.sql.SQLContext@73dbb666

scala> val df = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").option("inferSchema", "true").load("hdfs:///user/crimedata/crime.csv")
df: org.apache.spark.sql.DataFrame = [_c0: int, Date: string ... 7 more fields]

scala> df.printSchema()
root
 |-- _c0: integer (nullable = true)
 |-- Date: string (nullable = true)
 |-- Primary: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
```

Step 5: Delete fields containing null value

```

[scala> df.limit(20).show

|_c0_|_Date_|_Block_|_Primary_|_Description_|_Arrest_|_Year_|_Latitude_|_Longitude_| | |
|---|---|---|---|---|---|---|---|---|---|---|
|3|05/03/2016|11:40:...|013XX S SAWYER AVE|BATTERY|DOMESTIC BATTERY ...|true|2016|41.86407316|-87.70681861|
|89|05/03/2016|09:40:...|061XX S DREXEL AVE|BATTERY|DOMESTIC BATTERY ...|false|2016|41.78292153|-87.60436317|
|197|05/03/2016|11:31:...|053XX W CHICAGO AVE|PUBLIC PEACE|VIOL...|RECKLESS CONDUCT|false|2016|41.89490828|-87.75837196|
|673|05/03/2016|10:10:...|049XX W ELLIOT ST|BATTERY|SIMPLE|false|2016|41.88568651|-87.74951581|
|785|05/03/2016|10:01:...|null|BATTERY|DOMESTIC BATTERY ...|false|2016|null|null|
|911|05/03/2016|10:00:...|003XX N LOUIS AVE|THEFT|$500 AND UNDER|false|2016|41.88629742|-87.76175071|
|1108|05/03/2016|10:35:...|082XX S MARYLAND AVE|BATTERY|AGGRAVATED: HANDGUN|false|2016|41.74535402|-87.6037989|
|1130|05/03/2016|10:30:...|027XX S STATE ST|BATTERY|SIMPLE|false|2016|41.84402377|-87.62692325|
|1801|05/03/2016|09:30:...|002XX E 4TH ST|BATTERY|SIMPLE|false|2016|41.81113396|-87.62074077|
|1868|05/03/2016|04:00:...|014XX W DEVON AVE|BATTERY|SIMPLE|false|2016|41.99813061|-87.66581404|
|1891|05/03/2016|10:30:...|069XX S ASHLAND AVE|BATTERY|DOMESTIC BATTERY ...|false|2016|41.76809684|-87.6367859|
|1935|05/03/2016|09:44:...|074XX S SOUTH SHO...|WEAPONS VIOLATION|UNLAWFUL POSS OF ...|true|2016|41.76173329|-87.55830998|
|2150|05/03/2016|11:11:...|006XX N WABASH AVE|BATTERY|DOMESTIC BATTERY ...|true|2016|41.89302675|-87.62675083|
|2193|05/03/2016|05:30:...|011XX W JACKSON BLVD|THEFT|FROM BUILDING|false|2016|41.87781186|-87.65575801|
|2279|05/03/2016|09:00:...|028XX S DR MARTIN...|THEFT|$500 AND UNDER|false|2016|41.84301696|-87.61722727|
|2473|05/03/2016|10:08:...|016XX N CLAREMONT|THEFT|OVER $500|false|2016|41.81080883|-87.68618755|
|2580|05/03/2016|10:30:...|null|THEFT|$500 AND UNDER|false|2016|null|null|
|2847|05/03/2016|04:00:...|072XX S RICHMOND ST|BATTERY|DOMESTIC BATTERY ...|false|2016|41.76208943|-87.69663234|
|3023|05/03/2016|09:45:...|020XX W LE MOYNE ST|THEFT|OVER $500|false|2016|41.90823711|-87.67843742|
|3088|05/03/2016|10:00:...|055XX S STATE ST|ROBBERY|ARMED: OTHER DANG...|false|2016|41.79393591|-87.62568028|

```

```
scala> var res = df.na.drop()
res: org.apache.spark.sql.DataFrame = [_c0: int, Date: string ... 7 more fields]

scala> res.limit(20).show
```

_c0	Date	Block	Primary	Description	Arrest	Year	Latitude	Longitude
3	05/03/2016 11:40:...	013XX S SAWYER AVE	BATTERY	DOMESTIC BATTERY ...	true	2016	41.86407316	-87.70681861
89	05/03/2016 09:40:...	061XX S DREXEL AVE	BATTERY	DOMESTIC BATTERY ...	false	2016	41.78292153	-87.60436317
197	05/03/2016 11:31:...	053XX W CHICAGO AVE	PUBLIC PEACE VIOL...	RECKLESS CONDUCT	false	2016	41.89490828	-87.75837196
673	05/03/2016 10:10:...	049XX W FULTON ST	BATTERY	SIMPLE	false	2016	41.88568685	-87.74951598
911	05/03/2016 10:00:...	003XX N LOTUS AVE	THEFT	\$500 AND UNDER	false	2016	41.88629724	-87.76175071
1108	05/03/2016 10:35:...	082XX S MARYLAND AVE	BATTERY	AGGRAVATED: HANDGUN	false	2016	41.74535402	-87.6037989
1130	05/03/2016 10:30:...	027XX S STATE ST	BATTERY	SIMPLE	false	2016	41.84402377	-87.62692325
1801	05/03/2016 09:30:...	002XX E 46TH ST	BATTERY	SIMPLE	false	2016	41.81113396	-87.62074077
1868	05/03/2016 04:00:...	014XX W DEVON AVE	BATTERY	SIMPLE	false	2016	41.99813061	-87.66581404
1891	05/03/2016 10:30:...	069XX S ASHLAND AVE	BATTERY	DOMESTIC BATTERY ...	false	2016	41.76809684	-87.66387859
1935	05/03/2016 09:44:...	074XX S SOUTH SHO...	WEAPONS VIOLATION	UNLAWFUL POSS OF ...	true	2016	41.76173329	-87.55830998
2150	05/03/2016 11:11:...	006XX N WABASH AVE	BATTERY	DOMESTIC BATTERY ...	true	2016	41.89302675	-87.62675083
2193	05/03/2016 05:30:...	011XX W JACKSON BLVD	THEFT	FROM BUILDING	false	2016	41.87781196	-87.65575801
2279	05/03/2016 09:00:...	028XX S DR MARTIN...	THEFT	\$500 AND UNDER	false	2016	41.84301696	-87.61722727
2477	05/03/2016 10:08:...	016XX N CLAREMONT...	THEFT	OVER \$500	false	2016	41.91090083	-87.68601875
2847	05/03/2016 04:00:...	072XX S RICHMOND ST	BATTERY	DOMESTIC BATTERY ...	false	2016	41.76208913	-87.69663234
3023	05/03/2016 09:45:...	020XX W LE MOYNE ST	THEFT	OVER \$500	false	2016	41.9082371	-87.67843742
3088	05/03/2016 10:00:...	055XX S STATE ST	ROBBERY	ARMED: OTHER DANG...	false	2016	41.79393591	-87.62568028
3242	05/03/2016 08:00:...	100XX S SANGAMON ST	MOTOR VEHICLE THEFT	AUTOMOBILE	false	2016	41.71184357	-87.64660793

step 6: Save new CSV file to HDFS

```
scala> res.write.format("com.databricks.spark.csv").save("hdfs:///user/crimedata/newcrime.csv")
```

zhe_wang23@cluster-cd23-m: ~ - Google Chrome

ssh.cloud.google.com/projects/cloud-technology-assignment/zones/europe-west2-a/instances/cluster-cd23-m?authu

```
zhe_wang23@cluster-cd23-m:~$ hdfs dfs -ls /user/crimedata/
Found 2 items
-rw-r--r-- 2 zhe_wang23 hadoop 2302 2021-11-16 22:55 /user/crimedata/crime.csv
drwxr-xr-x - zhe_wang23 hadoop 0 2021-11-16 23:18 /user/crimedata/newcrime.csv
zhe_wang23@cluster-cd23-m:~$
```

Step 7 Export csv files from HDFS to google storage and download it to local couputer

```
zhe_wang23@cluster-cd23-m:~$ hadoop fs -cp /user/crimedata/newcrime.csv/ gs://hive-test1/cirme
zhe_wang23@cluster-cd23-m:~$
```

Buckets > hive-test1 > cirme		UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DOWNLOAD DELETE	
Filter by name prefix only		<input checked="" type="checkbox"/> Filter Filter objects and folders <input type="checkbox"/> Show deleted data	
Marketplace	<input checked="" type="checkbox"/>	Name	Size
Release notes	<input checked="" type="checkbox"/>	newcrime.csv/	Folder