Scala Spark Project

FINISHED

Annual Revenue Vs. Executive Pay

for

Recipients of U.S. Federal Funds

Data Source: USASpending.gov

According to their website: USAspending.gov is the publicly accessible, searchable website mandated by the Federal Funding Accountability and Transparency Act of 2006 to give the American public access to information on how their tax dollars are spent.

FINISHED

Data extraction

The data was downloaded as a batched CSV file from USASpending.gov. The CSV file contained a record of all federal funding allocation to companies and non-profits in 2015.

Of the 225 fields in the original data, we are most interested in dollarsobligated, vendorname, annualrevenue, prime_awardee_executive1_compensation, prime_awardee_executive2_compensation, prime_awardee_executive3_compensation, prime_awardee_executive4_compensation, and prime_awardee_executive5_compensation.

The original CSV file had 261 line breaks within individual data fields out of over 20M rows of data. The CSV could not be loaded into a Spark Dataframe with this inconsistent formatting. The problematic rows were removed from the data set for this exercise.

Here are the steps I took to clean the data from the UNIX command line:

1. Select for lines that begin and end with a quote.

   ```
   pv Data_Feed.csv | grep “^".*"\r” > data_feed_good.csv
   ```

2. Filter out lines that begin with an end-quote followed by a comma.

   ```
   pv data_feed_good.csv | grep -v “^",.*\r” > data_feed_good_2.csv
   ```

Note: These data cleansing steps remove 0.001305% of data.

Now, let's read the cleaned parquet file from S3 to a Spark data frame.

```
val df = sqlContext.read.parquet("s3://sarah-usaspendingfy2015/clean_data.parquet")
```

FINISHED

```
df: org.apache.spark.sql.DataFrame = [unique_transaction_id: string, transaction_status: st
ring ... 223 more fields]
```

```
df.columns
```

FINISHED

```
res235: Array[String] = Array(unique_transaction_id, transaction_status, dollarsobligated,
baseandexercisedoptionsvalue, baseandalloptionsvalue, maj_agency_cat, mod_agency, maj_fund_
agency_cat, contractingofficeagencyid, contractingofficeid, fundingrequestingagencyid, fund
ingrequestingofficeid, fundedbyforeignentity, signeddate, effectivedate, currentcompletiond
ate, ultimatecompletiondate, lastdatetoorder, contractactiontype, reasonformodification, ty
```

FINISHED

How many contracts, loans, and grants were obligated by the federal government in 2015?

```
df.count
```

FINISHED

```
res236: Long = 3679400
```

FINISHED

What companies have the most federal dollars obligated to them?

```
val dollarsObligated = df.select(df("vendorname"), df("dollarsobligated"))
    .groupBy("vendorname")
    .sum("dollarsobligated")
    .sort($"sum(dollarsobligated)".desc);
dollarsObligated.show()
```

FINISHED

```
dollarsObligated: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [vendorname: str
ing, sum(dollarsobligated): double]
+--------------------+---------------------+
|          vendorname|sum(dollarsobligated)|
+--------------------+---------------------+
|LOCKHEED MARTIN C...| 2.485973147905999...|
|    RAYTHEON COMPANY|  8.666833321250006E9|
| BOEING COMPANY, THE|  7.604776054149998E9|
|MCKESSON CORPORATION|  7.563494564229999E9|
|NORTHROP GRUMMAN ...|  6.322189713929999E9|
|ELECTRIC BOAT COR...|      5.02475097855E9|
|  THE BOEING COMPANY|      4.90264890895E9|
|SCIENCE APPLICATI...| 3.4890949293700027E9|
|HUMANA MILITARY H...|  3.313841091250001E9|
|HUNTINGTON INGALL...|      3.24046953884E9|
|SIKORSKY AIRCRAFT...| 3.1885837808500004E9|
|UNITED TECHNOLOGI...|      2.92949119238E9|
|L-3 COMMUNICATION...|      2.86854365592E9|
```

FINISHED

What companies have the highest self-reported annual revenue?

FINISHED

```
df.select(
    $"vendorname",
    $"annualrevenue")
    .sort( $"annualrevenue".desc)
    .distinct().show()
```

```
+--------------------+---------------+
|          vendorname|  annualrevenue|
+--------------------+---------------+
|SIERRA NEVADA COR...|937973411479552|
|   XEROX CORPORATION|179999999000576|
|GEOLOGICAL & ECON...|155529003401216|
|D & L WASTE EQUIP...| 99999999999999|
|          APPLE INC.| 46900000000000|
|          APPLE INC.| 46899998490624|
|METAL-FLEX WELDED...| 31582003920896|
|          APPLE INC.| 25600000000000|
|          APPLE INC.| 25599999475712|
|VERIZON BUSINESS ...| 25599999475712|
|       DRS ICAS, LLC| 25000000000000|
|BAE SYSTEMS LAND ...| 24613600000000|
|BAE SYSTEMS LAND ...| 24613599838208|
|      HEDGEROW FARMS| 21126761676800|
|          APPLE INC.| 19999999655936|
```

FINISHED

What companies have the highest executive compensation (as reported by max of executive compensation fields)?

The highest executive compensation reported is $950 Bilion by SCHAFER AEROSPACE.

```
import scala.util.Try
```

FINISHED

```
import scala.util.Try
```

```
val execCompensation = df.select(
    $"vendorname",
    $"prime_awardee_executive1_compensation",
    $"prime_awardee_executive2_compensation",
    $"prime_awardee_executive3_compensation",
    $"prime_awardee_executive4_compensation",
    $"prime_awardee_executive5_compensation")
    .rdd
    .map(row => {
        val vendorName: String = row.getString(0);
        val maxExecCompensation: Option[Double] =
            Try(Some((1 to 5).map(n =>
                Try(row.getDouble(n))
                .getOrElse(0.0)).max)).getOrElse(None);
        (vendorName, maxExecCompensation)
    })
    .toDF.sort(desc("_2")).distinct()
    .withColumnRenamed("_1", "vendorName")
    .withColumnRenamed("_2", "maxExecCompensation")
```

FINISHED

```
execCompensation: org.apache.spark.sql.DataFrame = [vendorName: string, maxExecCompensation
: double]
```

```
execCompensation.show()
```

FINISHED

```
+--------------------+-------------------+
|          vendorName|maxExecCompensation|
+--------------------+-------------------+
|SCHAFER AEROSPACE...|   9.53842289652E11|
| SCHAFER CORPORATION|   9.53842289652E11|
|RESEARCH ANALYSIS...|           1.704E11|
|RESEARCH ANALYSIS...|           1.704E11|
|FUTURE CARE HEALT...|       1.49900992E8|
|    ASE DIRECT, INC.|         1.180767E8|
|    A S E DIRECT LLC|         1.180767E8|
|PARAGON TECHNICAL...|           1.0658E8|
|DIXON GROUP, INC....|              1.0E8|
|         FORCE 3 INC|        8.8277633E7|
|       FORCE 3, INC.|        8.8277633E7|
|FORCE 3 INCORPORATED|        8.8277633E7|
|           ECCO GMBH|        7.4968679E7|
|ENVIRONMENTAL CHE...|        6.5707594E7|
|     AM GENERAL, LLC|        5.2454072E7|
```

FINISHED

What companies have the highest executive pay (as a percentage of total reported annual revenue)?

FINISHED

```
val percentExecComp = df.select(
    $"vendorname",
    $"prime_awardee_executive1_compensation",
    $"prime_awardee_executive2_compensation",
    $"prime_awardee_executive3_compensation",
    $"prime_awardee_executive4_compensation",
    $"prime_awardee_executive5_compensation",
    $"annualrevenue").rdd
    .map(row => {
        val vendorName: String = row.getString(0);
        val maxExecCompensation: Option[Double] =
            Try(Some((1 to 5).map(n =>
                Try(row.getDouble(n)).getOrElse(0.0)).max)).getOrElse(None);
        val annualRevenue: Option[Double] =
            Try(Some(row.getLong(6).toDouble)).getOrElse(None);
        val percentageCompensation: Option[Double] =
            Try(Some(maxExecCompensation.get/annualRevenue.get)).getOrElse(None);
        (vendorName, maxExecCompensation, annualRevenue, percentageCompensation)
    }).filter(row => row._3.exists(_ > 0))
    .distinct
    .toDF.sort($"_4".desc)
    .withColumnRenamed("_1", "vendorName")
    .withColumnRenamed("_2", "maxExecCompensation")
    .withColumnRenamed("_3", "annualRevenue")
    .withColumnRenamed("_4", "percent")
    .distinct
```

```
percentExecComp: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [vendorName: stri
ng, maxExecCompensation: double ... 2 more fields]
```

Some companies pay out high amounts to executives, while reported siginificanlty lower annual revenue. FINISHED

```
percentExecComp.show()
```

FINISHED

```
+--------------------+-------------------+-------------+--------------------+
|          vendorName|maxExecCompensation|annualRevenue|             percent|
+--------------------+-------------------+-------------+--------------------+
|GENERAL DYNAMICS ...|          1810115.0|    9000000.0| 0.20112388888888888|
|    PD SYSTEMS, INC.|           229058.0|    1765000.0| 0.12977790368271955|
|         KNWEBS INC.|           192000.0|    1950000.0| 0.09846153846153846|
|PROSOURCE CONSULT...|           445000.0|    5000000.0|               0.089|
|DANYA INTERNATION...|          5045608.0|  6.8066532E7| 0.07412759034058031|
|M & M CONTRACT MO...|            36000.0|     500000.0|               0.072|
|  J'S ASSOCIATES LLC|           163462.0|    2454904.0| 0.06658590315547981|
|LIFE SCIENCE LOGI...|           480000.0|    9726812.0| 0.04934813174141744|
|BRAINERD HELICOPT...|           220000.0|   1.163553E7|0.018907604552607402|
|WIEDEMANN CONSULTING|           240000.0|        1.5E7|               0.016|
|SCIENCE AND TECHN...|           809644.0|  5.2216524E7|0.015505513159014567|
|STRATEGIC RESOURC...|           257030.0|       1.85E7|0.013893513513513513|
|HERNDON PRODUCTS,...|           980914.0|  7.8336466E7| 0.01252180561732259|
|MATHEMATICA POLIC...|           766000.0|    6.60175E7|0.011602984057257546|
|    PRAGMATICS, INC.|           746459.0|       6.61E7|0.011292874432677761|
```

How many companies are there with annual revenue > $1M and executive compensation > $1M?

FINISHED

```
(percentExecComp.count, percentExecComp.filter($"maxExecCompensation" > 1000000).filter($"
```

FINISHED

```
res241: (Long, Long) = (196630,1458)
```

Of the companies with annual revenue > $1M and executive compensation > $1M, which companies pay out the highest amounts to executives?

FINISHED

```
percentExecComp.filter($"annualRevenue">=1000000).filter($"maxExecCompensation">=1000000).s
```

FINISHED

```
+--------------------+-------------------+---------------+--------------------+
|          vendorName|maxExecCompensation|  annualRevenue|             percent|
+--------------------+-------------------+---------------+--------------------+
|DANYA INTERNATION...|          5045608.0|    6.8066532E7| 0.07412759034058031|
|COX CONSTRUCTION CO.|          1147170.0|    3.0015833E7| 0.03821882937581642|
|PROPULSION CONTRO...|          2304231.0|    6.6175336E7| 0.03482008765320058|
|BAE SYSTEMS LAND ...|          3896785.0|3.0000001024E10|1.298928288996581E-4|
|GENERAL DYNAMICS ...|          1768749.0|3.1249000448E10|5.660177844546716E-5|
|GENERAL DYNAMICS ...|          1810115.0|      9000000.0| 0.20112388888888888|
|WORLDWIDE LANGUAG...|          1178000.0|         4.66E7|0.025278969957081544|
|DELL SERVICES FED...|          1050420.0|         2.75E8|0.003819709090909...|
|IRON BOW TECHNOLO...|          1377025.0|   7.25114408E8|0.001899045150403...|
|QINETIQ NORTH AME...|          1740126.0|          1.2E9|         0.001450105|
|NORTHROP GRUMMAN ...|        2.4411853E7| 3.213700096E10| 7.59618267752636E-4|
|ROLLS ROYCE NORTH...|          1271550.0|        2.284E9|5.567206654991244E-4|
|ENVIRONMENTAL CHE...|        4.9559414E7|         7.5E7|  0.6607921866666666|
|INDUSTRIAL ECONOM...|          4224288.0|     3.06666E7| 0.13774882119308954|
|COX CONSTRUCTION CO.|          1074897.0|    2.9234088E7| 0.03676861751254221|
```

FINISHED

What companies have the highest executive pay (as a percentage of sum dollars obligated)?

FINISHED

```
val joined_df = dollarsObligated.toDF.join(
    execCompensation.toDF, "vendorName");


joined_df.show()
```

```
+--------------------+---------------------+-------------------+
|          vendorname|sum(dollarsobligated)|maxExecCompensation|
+--------------------+---------------------+-------------------+
|SCHAFER AEROSPACE...|              50000.0|   9.53842289652E11|
| SCHAFER CORPORATION|  5.523365843000001E7|   9.53842289652E11|
|RESEARCH ANALYSIS...|                  0.0|           1.704E11|
|RESEARCH ANALYSIS...|           -2285908.6|           1.704E11|
|FUTURE CARE HEALT...|    595829.1399999999|       1.49900992E8|
|    ASE DIRECT, INC.|    6241757.649999998|         1.180767E8|
|    A S E DIRECT LLC|   2009835.9099999997|         1.180767E8|
|PARAGON TECHNICAL...| 1.0600045000000002E7|           1.0658E8|
|DIXON GROUP, INC....|            -33444.09|              1.0E8|
|         FORCE 3 INC|        1.111120801E7|        8.8277633E7|
|       FORCE 3, INC.|  6.932773838999999E7|        8.8277633E7|
|FORCE 3 INCORPORATED|           6477692.14|        8.8277633E7|
|           ECCO GMBH|    5067407.069999999|        7.4968679E7|
|ENVIRONMENTAL CHE...| 4.289280983000006E7|        6.5707594E7|
|     AM GENERAL, LLC|  5.318833116000001E7|        5.2454072E7|
```

FINISHED

maxExecCompensation is often reported as 0.0 or 1.0.

```
joined_df.select($"vendorname",
                 $"maxExecCompensation",
                 $"sum(dollarsobligated)",
                 ($"sum(dollarsobligated)" / $"maxExecCompensation").as("percent"))
                 .sort($"percent".desc).show()
```

FINISHED

```
+--------------------+-------------------+---------------------+--------------------+
|          vendorname|maxExecCompensation|sum(dollarsobligated)|             percent|
+--------------------+-------------------+---------------------+--------------------+
|LAWRENCE LIVERMOR...|                1.0|      1.42804417076E9|     1.42804417076E9|
|COUNTERTRADE PROD...|                1.0| 1.5911857584999987E8|1.5911857584999987E8|
|GOODRICH CORPORATION|                1.0|        1.3467258139E8|      1.3467258139E8|
|CSI AVIATION SERV...|                1.0|  9.029598071000001E7| 9.029598071000001E7|
|WILDFLOWER INTERN...|                1.0|  5.143629600000001E7| 5.143629600000001E7|
|SIEMENS GOVERNMEN...|                1.0|         3.423444534E7|       3.423444534E7|
|NCR GOVERNMENT SY...|                1.0| 2.8042943869999997E7|2.8042943869999997E7|
|SIMMONDS PRECISIO...|                1.0| 2.4675061740000002E7|2.4675061740000002E7|
|      HUMANTOUCH LLC|                1.0| 2.1046803720000003E7|2.1046803720000003E7|
|SUPERTEL NETWORK ...|                1.0|           5805720.44|          5805720.44|
|PEARSON ENGINEERI...|                1.0|           5600751.34|          5600751.34|
|TATITLEK CONSTRUC...|                1.0|    4530614.670000001|   4530614.670000001|
|PROTECTION STRATE...|                1.0|    4276807.710000001|   4276807.710000001|
|CARLETON TECHNOLO...|                1.0|           3964837.21|          3964837.21|
|SYLVAIN ANALYTICS   |                1.0|           3919515.84|          3919515.84|
```

How many companies have reported executive compensation?

FINISHED

```
(joined_df.count, joined_df.filter($"maxExecCompensation" > 0).count)
```

FINISHED

```
res245: (Long, Long) = (154596,4536)
```

How many companies have reported executive compensation over $1M?

FINISHED

```
(joined_df.count, joined_df.filter($"maxExecCompensation" >= 1000000).count)
```

FINISHED

```
res246: (Long, Long) = (154596,799)
```

Of the companies that pay executives more than $1M, which have the highest

FINISHED

executive pay (as a percentage of sum dollars obligated)?

FINISHED

```
joined_df.select($"vendorname",
                 $"maxExecCompensation",
                 $"sum(dollarsobligated)",
                 ($"sum(dollarsobligated)" / $"maxExecCompensation").as("percent"))
                 .filter($"maxExecCompensation" >= 1000000)
                 .sort($"percent".desc).show()
```

```
+--------------------+-------------------+---------------------+------------------+
|          vendorname|maxExecCompensation|sum(dollarsobligated)|           percent|
+--------------------+-------------------+---------------------+------------------+
|ELECTRIC BOAT COR...|          1334063.0|  5.024750978549999E9| 3766.502015684416|
|ELECTRIC BOAT COR...|          1919901.0|  5.024750978549999E9|2617.1927503293136|
|ELECTRIC BOAT COR...|          2010674.0|  5.024750978549999E9| 2499.038122813544|
|         SANDIA CORP|          1337700.0| 2.8087110097799997E9|2099.6568810495623|
|ELECTRIC BOAT COR...|          2533838.0|  5.024750978549999E9|1983.0592873538085|
|LOS ALAMOS NATION...|          1082776.0| 2.0885016060100002E9| 1928.839950285193|
|LOS ALAMOS NATION...|          1174339.0| 2.0885016060100002E9|1778.4486472900928|
|UNITEDHEALTH MILI...|          1461163.0| 2.5053445468599997E9|1714.6235887850976|
|HEALTH NET FEDERA...|          2157971.0|      2.78254386023E9|1289.4259747837204|
|HEALTH NET FEDERA...|          2359163.0|      2.78254386023E9|1179.4623178771453|
|HEALTH NET FEDERA...|          2382560.0|      2.78254386023E9|1167.8798688091802|
|UNITED LAUNCH SER...|          1971781.0| 2.0998831267800002E9|1064.9677255131276|
|UNITED LAUNCH SER...|          2098482.0| 2.0998831267800002E9|1000.6676858700719|
|GENERAL ATOMICS A...|          1794950.0| 1.7455848524600003E9| 972.4977589682165|
|UNITED LAUNCH SER...|          2254377.0| 2.0998831267800002E9|  931.469371263103|
```

FINISHED

Further Analysis

- Need to deal with duplicate and differing vendor information.
- Could some ML approach be applied to match vendors like “BOEING COMPANY, THE” and “THE BOEING COMPANY”.

FINISHED