Harvard University

CSCI-E88C | EUMAR ASSIS

Analyzing the impact of lobbying in the U.S through big data technologies

Project’s Goals

Use big data technologies to understand the impact of lobbying in U.S. politics. By reviewing an open lobbying dataset, we seek to investigate the flow of money in politics and provide analytics to strengthen democracy.

Drive Transparency

WE THE PEOPLE
COUNT MY VOICE COUNT MY VOTE
#AIGAvote #GetOutTheVote #ardesignshop #designfordemocracy

Strengthen Democracy

Track Flow of Money

How

1. Leverage the lobbying dataset provided by OpenSecrets.org, a nonpartisan, independent and nonprofit, premier research group tracking money in U.S. politics and its effect on elections and public policy.
   - Free for research and non-commercial purposes.
   - More than five million records.
   - 2GB in size.
2. Use **Scala** and **Spark** to analyze OpenSecrets' Lobbying dataset providing answers to fundamental questions that identify the impact of lobby in U.S. politics:

Total Amount of Money Spent in Lobbying

Top Lobbyist Organizations

Top Lobbied Bills

Top Industries Receiving Lobby

Former Officials Who Are Lobbyists

Solution Design

Scala Application

- Download OpenSecrets Dataset (CSV Files)
- Read CSV files defined in config file
- Use Spark SQL APIs in Scala to perform aggregations
- Save aggregations to materialized view (Hive tables)
- Unit Testing

Cloud Distributed Deployment

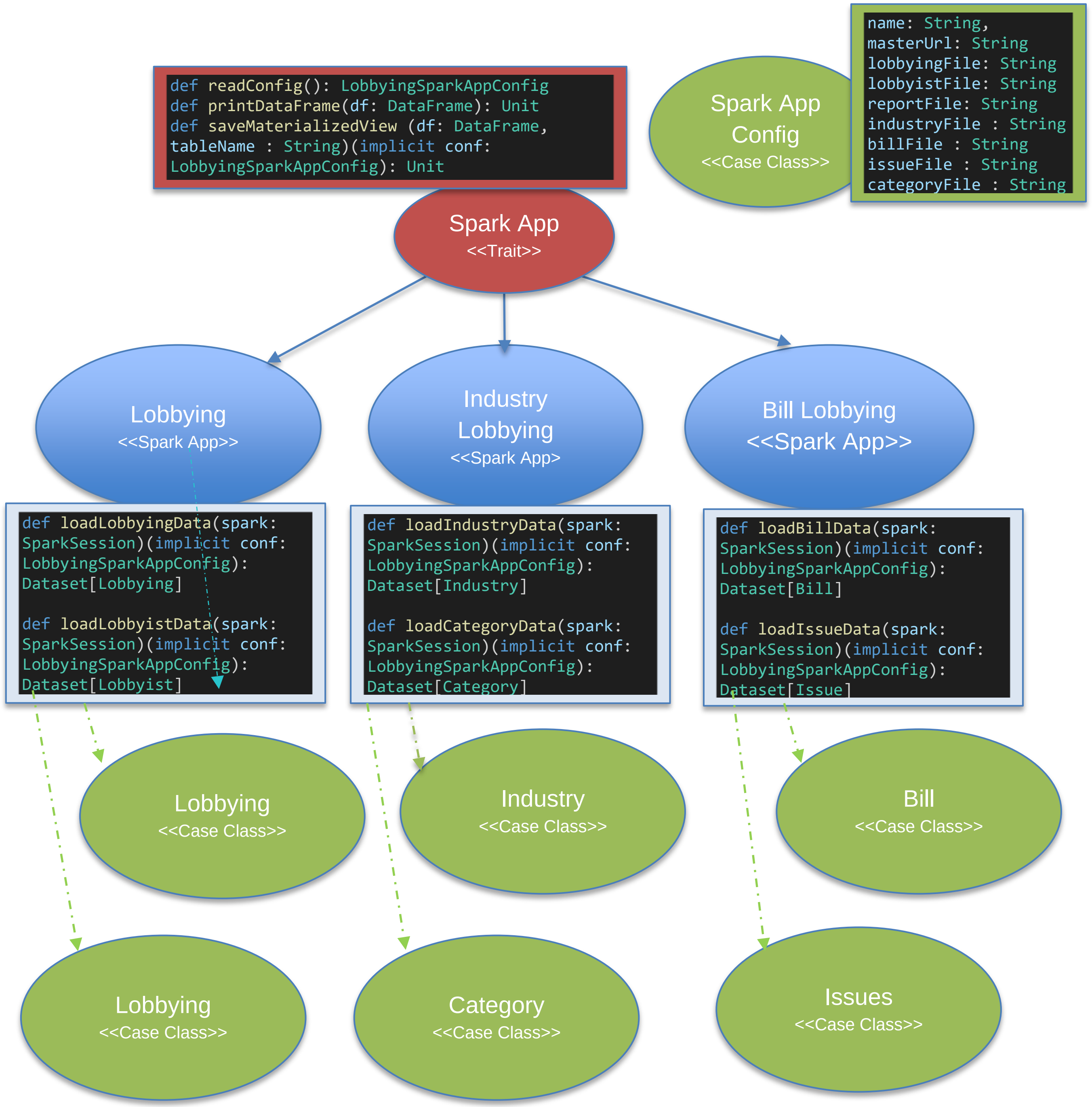
- CSV files deployed to Microsoft Azure Storage Account
- Scala App deployed to Azure Databricks cluster (Spark) as Job
- Materialized Summary Views on Hive Tables

Visualization

- Databricks Notebook with Built-in Visualizations
- Read Data from Materialized Summary Views on Hive Tables

Scala Implementation

System Design



```
name: String,
masterUrl: String
lobbyingFile: String
lobbyistFile: String
reportFile: String
industryFile : String
billFile : String
issueFile : String
categoryFile : String
```

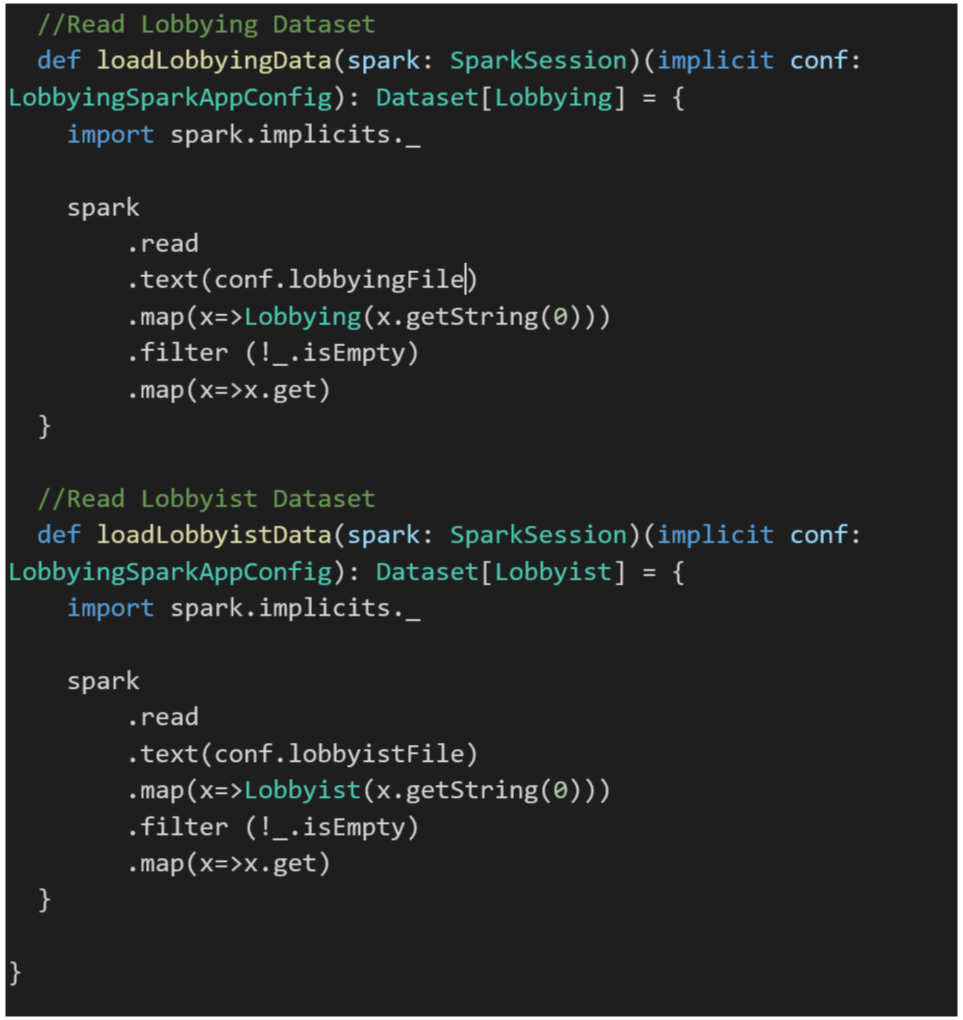
```
def readConfig(): LobbyingSparkAppConfig
def printDataFrame(df: DataFrame): Unit
def saveMaterializedView (df: DataFrame,
tableName : String)(implicit conf:
LobbyingSparkAppConfig): Unit
```

```
def loadLobbyingData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Lobbying]

def loadLobbyistData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Lobbyist]
```

```
def loadIndustryData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Industry]

def loadCategoryData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Category]
```

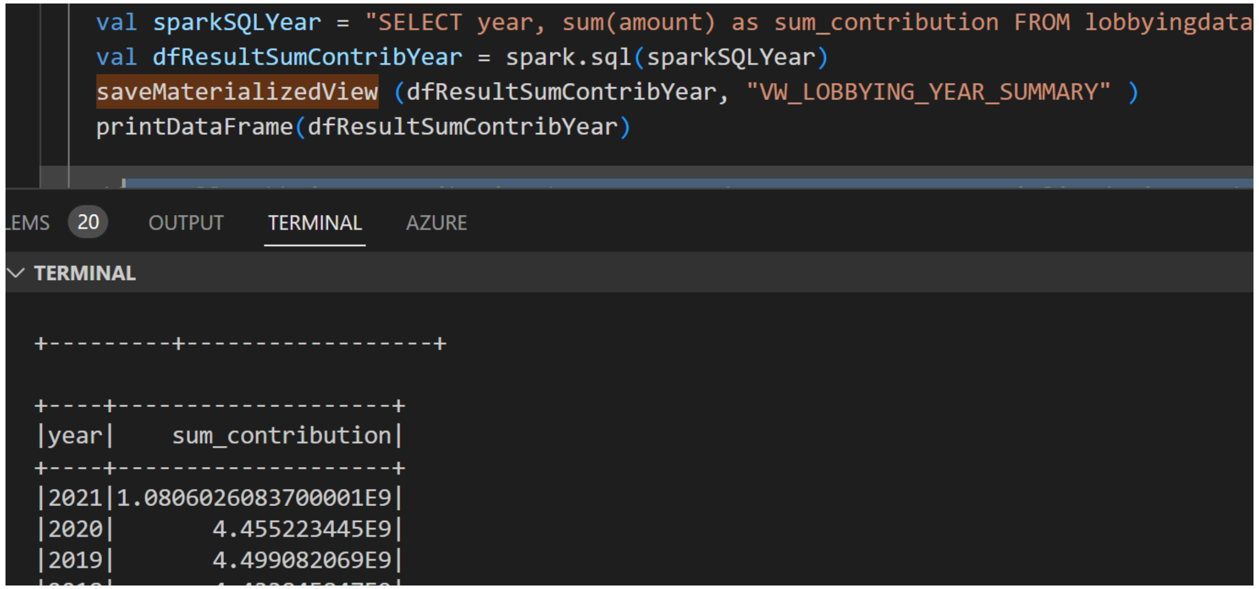
```
def loadBillData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Bill]

def loadIssueData(spark:
SparkSession)(implicit conf:
LobbyingSparkAppConfig):
Dataset[Issue]
```

Scala: Reading Dataset from Files



```
  //Read Lobbying Dataset
  def loadLobbyingData(spark: SparkSession)(implicit conf:
LobbyingSparkAppConfig): Dataset[Lobbying] = {
    import spark.implicits._

    spark
        .read
        .text(conf.lobbyingFile)
        .map(x=>Lobbying(x.getString(0)))
        .filter (!_.isEmpty)
        .map(x=>x.get)
  }

  //Read Lobbyist Dataset
  def loadLobbyistData(spark: SparkSession)(implicit conf:
LobbyingSparkAppConfig): Dataset[Lobbyist] = {
    import spark.implicits._

    spark
        .read
        .text(conf.lobbyistFile)
        .map(x=>Lobbyist(x.getString(0)))
        .filter (!_.isEmpty)
        .map(x=>x.get)
  }

}
```

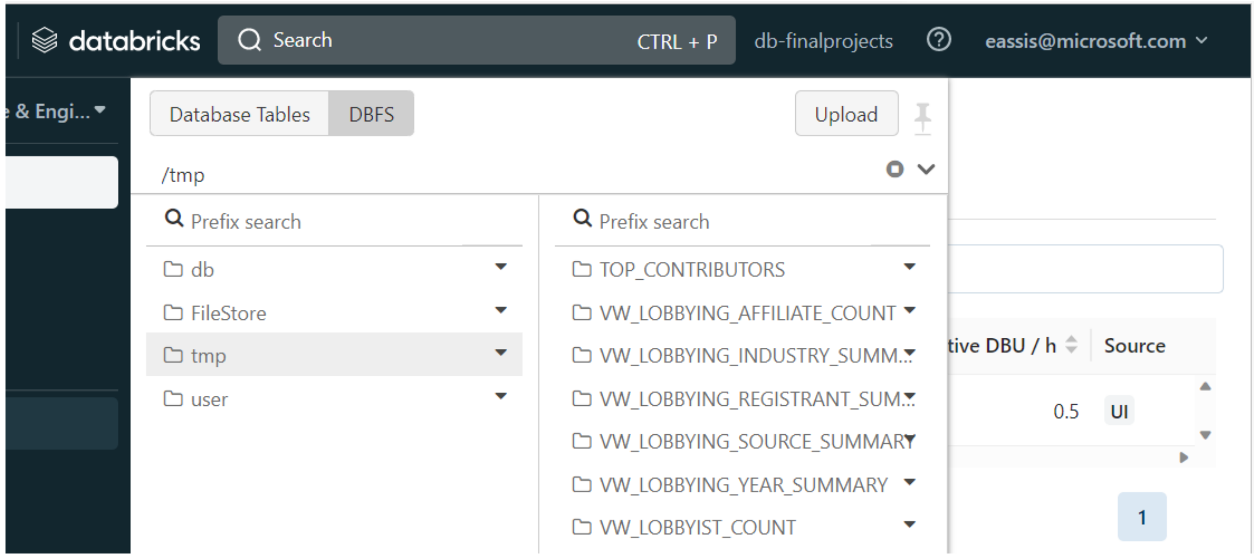
Scala: Aggregation using Spark SQL



```
val sparkSQLYear = "SELECT year, sum(amount) as sum_contribution FROM lobbyingdata
val dfResultSumContribYear = spark.sql(sparkSQLYear)
saveMaterializedView (dfResultSumContribYear, "VW_LOBBYING_YEAR_SUMMARY" )
printDataFrame(dfResultSumContribYear)
```

```
+---------+-----------------+

+----+-------------------+
|year|   sum_contribution|
+----+-------------------+
|2021|1.0806026083700001E9|
|2020|        4.455223445E9|
|2019|        4.499082069E9|
```

Databricks: Materialized Views

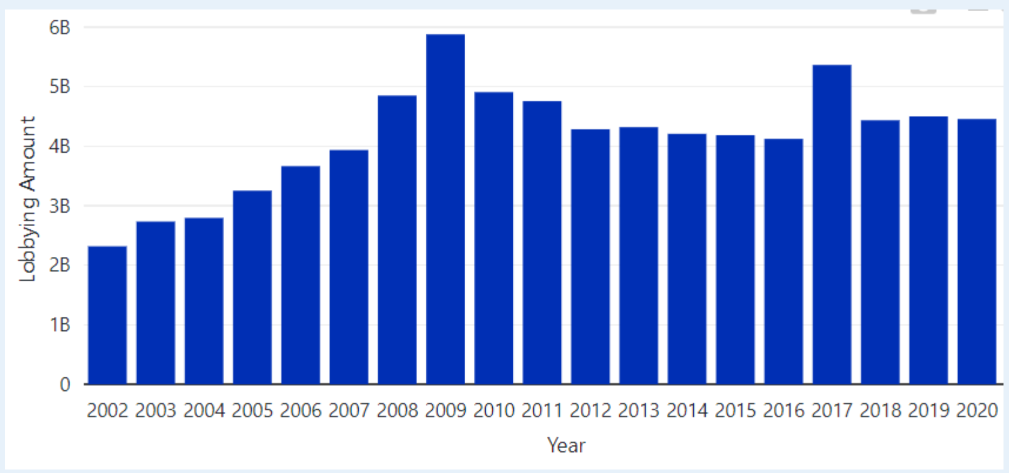


/tmp

Prefix search	Prefix search
db	TOP_CONTRIBUTORS
FileStore	VW_LOBBYING_AFFILIATE_COUNT
tmp	VW_LOBBYING_INDUSTRY_SUMM...
user	VW_LOBBYING_REGISTRANT_SUM...
	VW_LOBBYING_SOURCE_SUMMARY
	VW_LOBBYING_YEAR_SUMMARY
	VW_LOBBYIST_COUNT

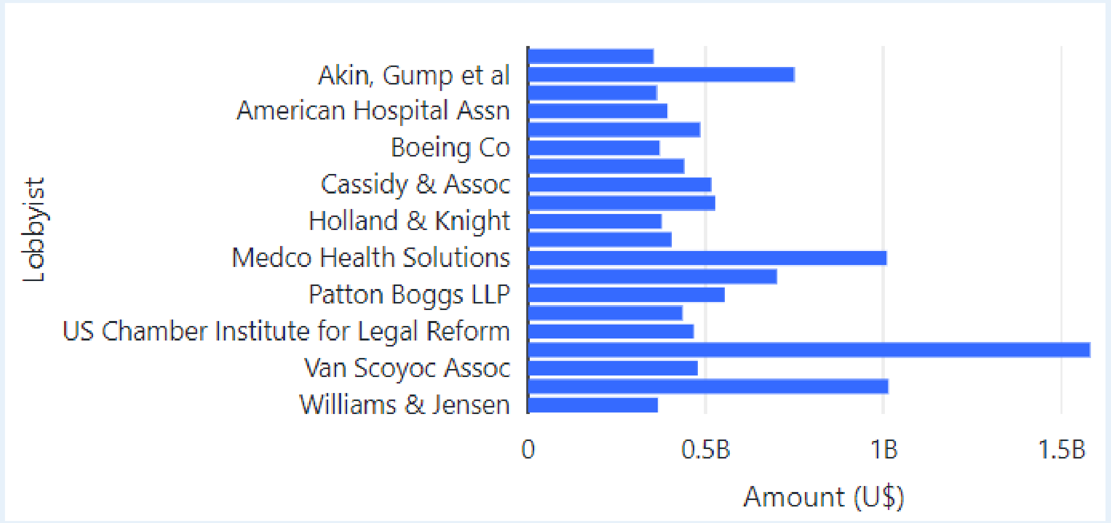
Results: U.S. Lobbying Analytics

Lobbying Amount per Year

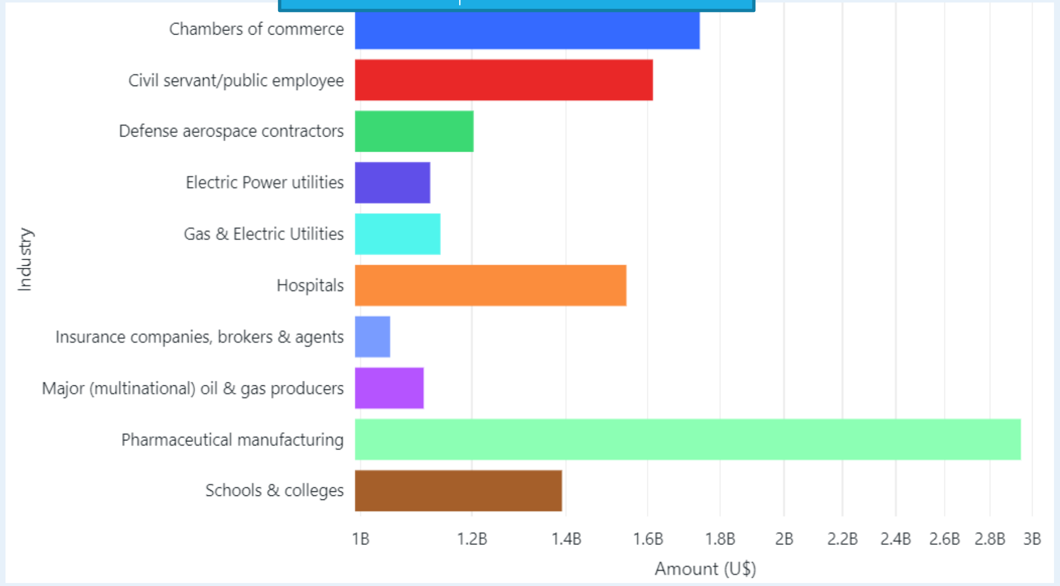


Lobbying Word Count

Top Lobbyists



Top Industries


Chambers of commerce
Civil servant/public employee
Defense aerospace contractors
Electric Power utilities
Gas & Electric Utilities
Hospitals
Insurance companies, brokers & agents
Major (multinational) oil & gas producers
Pharmaceutical manufacturing
Schools & colleges
Industry
1B
1.2B
1.4B
1.6B
1.8B
2B
2.2B
2.4B
2.6B
2.8B
3B
Amount (U$)