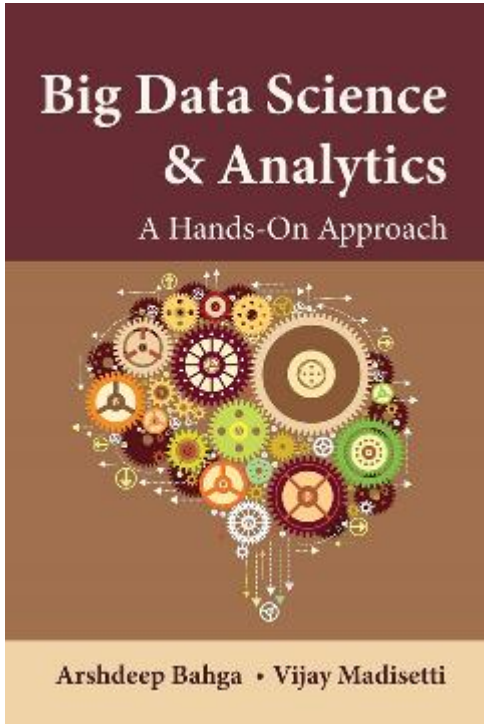
# Interactive Querying

**Prof. Gheith Abandah**

# Reference

- Chapter 9: **Interactive Querying**




- Arshdeep Bahga and Vijay Madisetti, **Big Data Science and Analytics: A Hands-On Approach**, 2019.
  - Web site: http://www.hands-on-books-series.com/

# Introduction

- Interactive querying is useful when your analytics application demands **flexibility to query data on demand**.
- **Common Framework**
  - Spark SQL
  - Hive
  - Google BigQuery
  - Amazon RedShift
- These frameworks allow users to **query data by writing statements** in **SQL-like** languages.

# Spark SQL

- Can interactively query **structured** and **semi-structured** data using **SQL-like** queries.
- Uses **DataFrames** which are distributed collections of data organized into named columns.
- DataFrames can be created from **existing RDDs**, **structured data files** (such as text files, JSON), and from **external databases**.
- To **launch** the Spark Python shell

```
bin/pyspark
```

# Spark SQL Example

- **SQLContext** is the entry point for Spark SQL.
- A DataFrame can be created by specifying the **RDD** and **schema**.

```
from pyspark.sql import SQLContext, Row
sqlContext = SQLContext(sc)
from pyspark.sql.types import *

lines =
  sc.textFile("file:///home/hadoop/file.csv")
parts = lines.map(lambda l: l.split(","))

ngrams = parts.map(lambda x: (x[0], x[1],
  x[2], x[3], x[4]))
schemaString = "ngram year count pages books"
fields = [StructField(field_name,
  StringType(), True) for field_name in
  schemaString.split()]
schema = StructType(fields)
schemaNGrams =
  sqlContext.createDataFrame(ngrams, schema)
```

# Spark SQL Example

- Spark SQL DataFrames support many **useful methods**
  - `show()`
  - `printSchema()`
  - `filter()`
  - `groupBy()`

```
schemaNGrams.groupBy("year").count().show()

+----+-----+
|year|count|
+----+-----+
|1831|   79|
|1832|   57|
…
+----+-----+
```

# Spark SQL Example

- Spark SQL allows **registering a DataFrame** as a temporary **table** for querying the data using **SQL-like queries**.

```
schemaNGrams.registerTempTable("ngrams")

result = sqlContext.sql("SELECT ngram,
  count FROM ngrams WHERE count
  >= 5").show()

+------+------+
| ngram| count|
+------+------+
|  ! 09|    17|
|! 1944|     8|
|  ! 28|    15|
…
+------+------+
```