Graphical Domain Specific Language For Solving Small and Medium Data Science Problems

Ahmed Issawi

Outline

- ❑ What is Data Science ?
- ❑ The Data Science Process ?
- ❑ Why should we use or implement DSL to solve the data science problems ?
- ❑ Graphical DSL For Solving Data Science Problems
- ❑ Case Study (Analyzing Application Server Log Files).

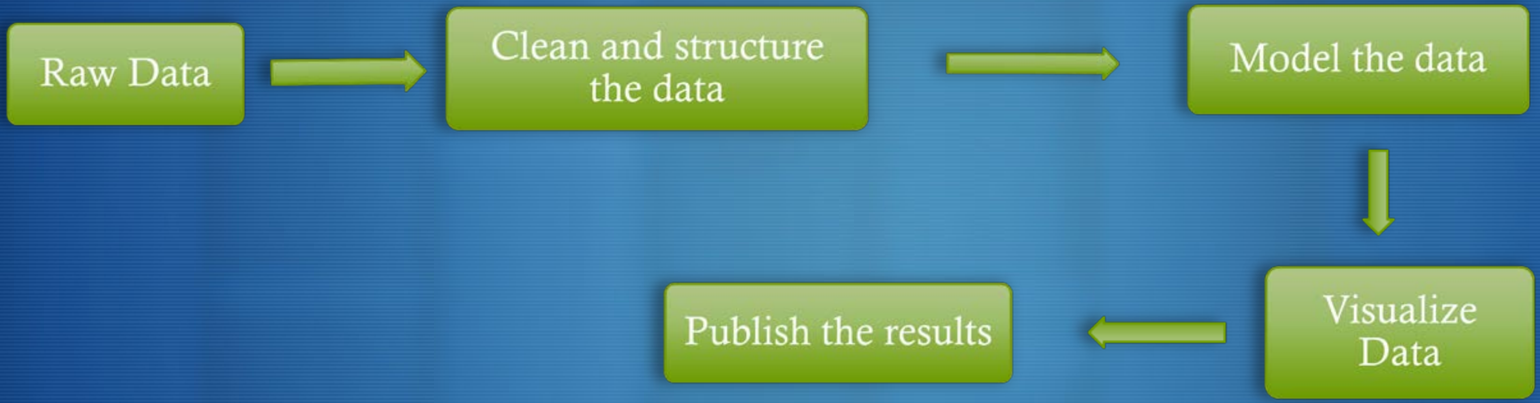
What is Data Science ?

- Data Science is based on designing and implementing data products from cleaning and munging the data to visualize it in the format that can make many users understand it easily.

- Data Science uses many existing sciences and concepts like computer science , statistics and machine learning.

- The Data scientist should know statistics more than any software engineer and knows software development more than any statistician.

The Data Science Process



Why should we use or implement DSL to solve the data science problems ?

- In this time we can see many programming languages like R ,Python Data Analysis Tools and Julia that have many libraries and extensions that can help us to solve those kind of problems , so why should we need DSL?

- DSL can give us the ability to concentrate on our problem domain , so we can solve complex problems faster and more accurate.

Why should we use or implement DSL to solve the data science problems ?

- If we are in company that perform data analysis methods and have experienced users that use some software tools , DSL will give them the flexibility to design permanent solutions to their daily problems easily.

- Although they have experience in their domain , they don’t have enough knowledge about software development or about big data algorithms.

Graphical DSL For Solving Data Science Problems

- DSL could give the experienced users the ability to design their solutions without a deep knowledge about the internal technical components.

- So if we mention the experienced users , why it is graphical DSL not textual one?

- Because we want to make the normal users be able to use our language like the experienced users.

Case Study

- Let’s take a look on scenario that can describe our proposed solution .
- Johns and Hana are working in software company .They are using monitoring tool to monitor their application servers but unfortunately the tool can’t process huge log files that are generated over months or years , so they are not be able to predict the downtime likelihood of their application servers , so they decided to build their tool. They don’t have enough experience to solve the big data tools that use systems like Hadoop ‘s ecosystem , moreover they don’t know much about the machine learning algorithms.
- They were searching in the web for any tool or even articles that can help them solving their problems.

Case Study

- They just found separate tools .Every tool can help them to solve only one part of their problem.

- In this scenario our language can help them. They will use it to store the big data by the help of Hadoop , they can print some exploratory data analysis ,explore the data patterns, apply some machine learning algorithms and predict upcoming events by using inference data analysis, visualize the data in nice graphs and publish their results in well organized web pages.

- The user couldn’t be aware about the internal structure of the language and its internal components.

- They just design the solution and our language could generate the suitable outputs

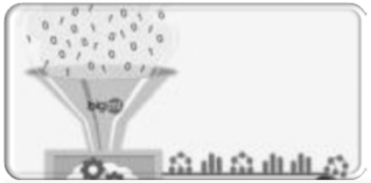
Case Study(How The Users Can Solve Their Problem By Using Our Language)

Raw Data

- The user should enter the location of the application server log files
- He should enter the required parameters like what he is looking for in this data and what kind of algorithms he wants apply on the data.

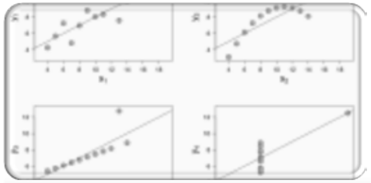
Clean and structure the data

- The system will store the data on distributed file store system like Hadoop File System.
- It will try convert the unstructured data to tidy data.
- It will run some Text Mining algorithms using Map Reduce Algorithms to search for the common words that are written on the log files .
- The user will be able to run some Exploratory Data Analysis.



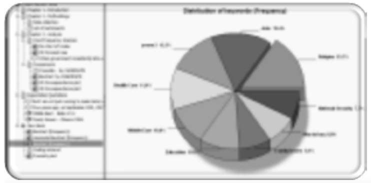
Model the data

- The system will try to find and run the suitable statistical models that can fit to the problem.
- It will find and run the machine learning algorithms that can fit.
- It will run the statistical inference data analysis methods on the data.



Visualize Data

- The system will generate the suitable graphs .



Publish the results

- The system will generate the analysis result
- The user will be able to view the final report of the data science process.