Operational Data Science

Posted by **datumengineering** on December 12, 2013

The term “Data Science” has been evolving not only as a niche skill but as a niche process as well. It is interesting to study “how” the Big data analytics/Data Science/Analytics can be efficiently implemented into the enterprise. So, along with my typical study of analytics viz. Big data analytics I have been also exploring the methodologies to bring the term “Data Science” into mainstream of existing enterprise data analysis, which we conventionally know as “Datawarehouse & BI”. This excerpt is just a study of Data Science workflow with respect to enterprise and opens the forum for discussion on Operational Data Science” (I am just tossing this term “Operational Data Science”, it can be named better!). Meanwhile, I must mention the articles those I have followed during my whole course of learning on the operational side of Data Science . Both the articles mentioned below are super write ups written by Data scientists during their research work and they can prove to be a valuable gift for enterprises.

1. **Data Science Workflow: Overview & Challenges.**
2. **Enterprise Data Analysis and Visualization: An Interview Study**

While article #1 gives fair idea of complete data science workflow, it can be very well understood by article #2 with nice explanation and challenges mentioned. Idea is to understand them together.

All the studies around implementation of analytics revolves around 5 basic areas: 1) Data Discovery 2) Data Preparation & Cleansing 3) Data Profiling 4) Modeling 5) Visualization.

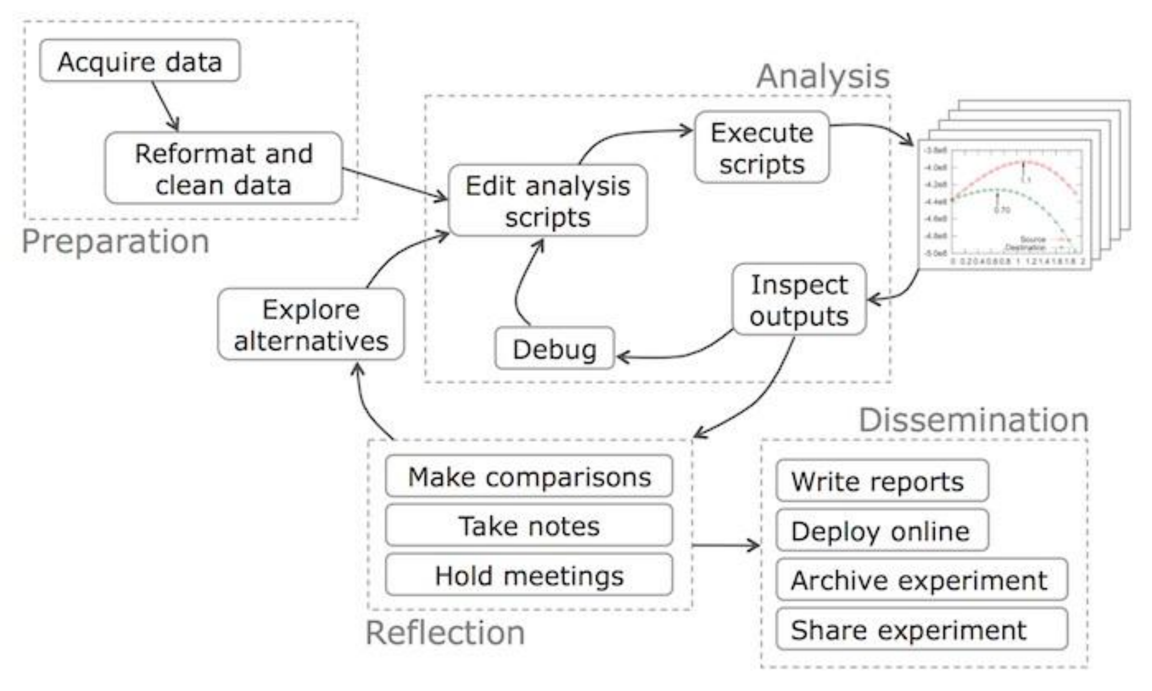
Operationally, these 5 areas can be efficiently covered if data scientist can rightly collaborate with Data Engineers, Datawarehouse Architects & Data Analyst. It is the responsibility of a Data Scientist to run the show from data discovery to communicating predictions to the business. I certainly don’t intend to define the role of a data scientist here (In fact i am not even eligible for this). My aim is to sum up the skill sets and identify operational aspect of it. One of the imprtant point to be discussed here is ‘Diversity of skills’.

Diversity is pretty important. A generalist is more valuable than a specialist. A specialist isn’t fluid enough.
We look
for pretty broad skills and data passion. If you are passionate about it you’ll jump into whatever tool you
need to. If
it’s in X, I’ll go jump in X.

It needs diversity of the skill i.e. “Business understanding on data (data discovery)” to “Programming hand on the scripting or SQL” to “communicate effectively through right visualization tool”. It is difficult for one person to diversify in all these areas and same time specialize in one. In such complex environment we should look at the opportunity to bring Datawarehouse + Unstructured Data analysis + Predictive Analytics together. This opportunity is well detailed in the article#2.

Most of the organisations work in silos on their data and in absence of effective communication channel between Datawarehouse and Analytics team the whole effort of effective analysis goes awry. 80% organisations divide the effort of Datawarehouse, Advance analytics & Statistical analysis into different teams and these teams not only address the different business problems but they also aligned with different architects. In my opinion this could be the main reason that kills the flavor of Data Science. Interestingly during one of my assignments in the field of retail data analysis, I observed that they had developed their datawarehouse team only at the maturity level of summarization and aggregation. I realized that this datawarehouse or Data store world would end after delivering bunch of reports and some dashboards. Eventually, they would be more interested in archiving the data thereafter. That's the Bottleneck !

To overcome this bottleneck we need to bring analytics either into mainstream of data processing layer or we should develop parallel workflow for this, and article #1 articulates the same and proposes the flow mentioned below. If you believe this figure (from article #1) is a data science workflow then you need to have diverse skilled engineers working on common goal to deliver this workflow unlike conventional data analysis. Observe it closely and figure out the business oriented engineering team.

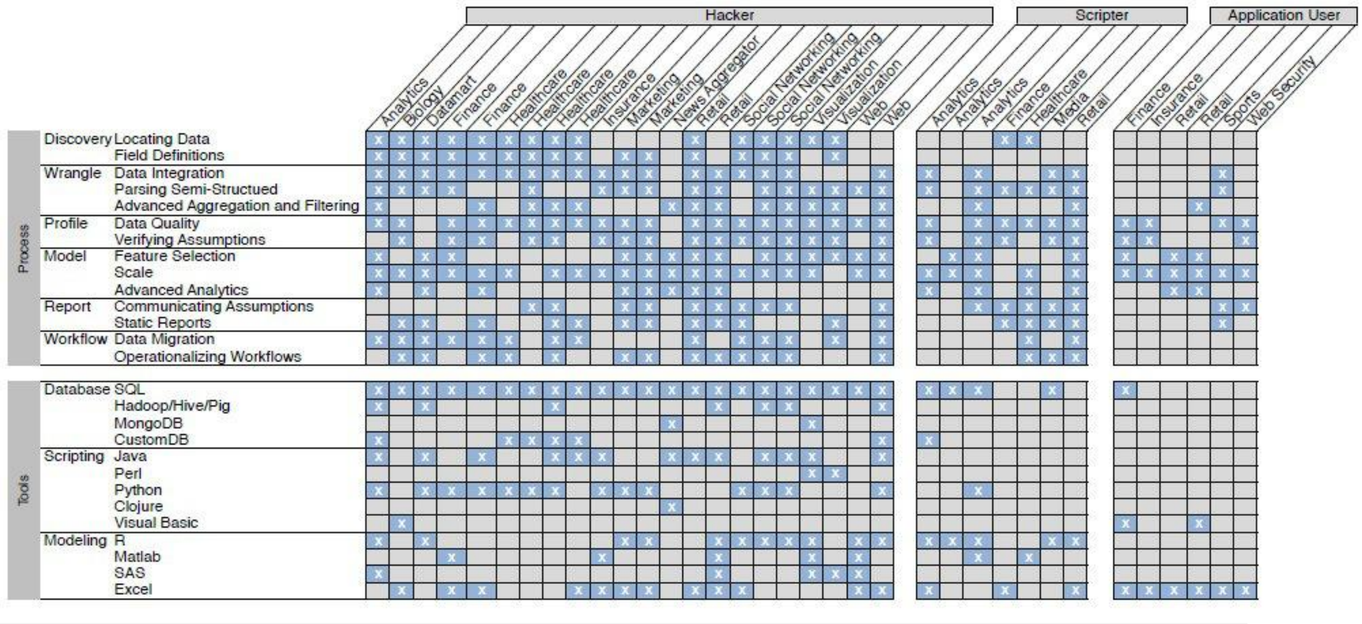


Team with diversity can work together for an enterprise to deliver this workflow.

- Person(s) with strong business acumen with Visualization skills.

- Data Integration Engineer(s) with ETL skills on structured and unstructured data
- Statistician with R skills
- Smart programmer with Predictive modeling skills.

Interestingly, there is no right and specific order of delivery from these people. Having said that the person who has strong business background can work at both the ends of a shore i.e. in data discovery as well as in communicating the final result (either in terms of prediction or pattern or summarization). However, programmers can pretty much independently work in all areas of data preparation, data analysis and scripting to build datasets for modeling (In fact this is hardest area read the article #2). In a same way, statistician can very much communicate the business result and reflection. Now after all these efforts what left is just a game of effective collaboration. That is easily visible in the figure mentioned below.



			Hacker																						Scripter							Application User					
			Analytics	Biology	Datamart	Finance	Finance	Healthcare	Healthcare	Healthcare	Healthcare	Insurance	Marketing	Marketing	News Aggregator	Retail	Retail	Social Networking	Social Networking	Social Networking	Visualization	Visualization	Web	Web	Analytics	Analytics	Analytics	Finance	Healthcare	Media	Retail	Finance	Insurance	Retail	Retail	Sports	Web Security
Process	Discovery	Locating Data	X	X	X	X	X	X	X	X	X					X		X	X	X	X	X						X	X								
		Field Definitions	X	X	X	X	X	X	X	X	X		X	X		X		X	X	X		X															
	Wrangle	Data Integration	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X				X	X		X			X	X					X	
		Parsing Semi-Structued	X	X	X	X			X			X	X	X		X	X		X	X	X	X	X	X	X		X	X	X	X	X					X	
		Advanced Aggregation and Filtering	X				X		X	X	X				X	X	X		X	X	X	X		X			X				X				X		
	Profile	Data Quality	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X			X	X
		Verifying Assumptions		X		X	X		X	X		X	X	X		X	X	X	X	X	X	X		X	X		X	X		X	X	X	X				X
	Model	Feature Selection	X		X	X							X	X	X	X	X		X	X	X	X	X	X		X	X				X	X		X	X		
		Scale	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X		X		X	X	X	X	X	X	X
		Advanced Analytics	X		X		X						X	X	X	X	X								X		X		X		X			X	X		
	Report	Communicating Assumptions							X	X			X	X		X	X	X	X	X				X			X	X	X	X	X					X	X
		Static Reports		X	X		X			X	X		X	X		X	X	X				X		X				X	X	X	X					X	
	Workflow	Data Migration	X	X	X	X	X	X		X	X					X		X	X	X		X		X					X		X						
		Operationalizing Workflows		X	X		X	X		X			X	X		X	X	X	X	X				X					X	X	X						
Tools	Database	SQL	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X		X					
		Hadoop/Hive/Pig	X		X					X							X		X	X				X													
		MongoDB													X						X																
		CustomDB	X					X	X	X	X													X	X												
	Scripting	Java	X		X		X			X	X	X			X	X	X		X	X	X			X													
		Perl																			X	X															
		Python	X		X	X	X	X	X	X		X	X	X				X	X	X				X			X										
		Clojure													X																						
		Visual Basic		X																												X			X		
	Modeling	R	X		X								X	X			X	X	X	X	X		X	X	X	X	X			X	X						
		Matlab				X						X					X				X		X				X		X								
		SAS	X														X				X	X	X														
		Excel		X		X	X				X	X	X	X		X	X	X					X	X	X			X			X	X	X	X	X	X	X

Responsibility Matrix

Moreover, along with the right collaboration channel there should be a Data scientist(s) who can watch over and architect the whole work flow and should always be ready to design+code+test the prototype of the end product. So, this whole Operation Data Science need a collaborative team and an architect(s) with diverse skills who should be ready to phrase the below statement.

Source: http://datumengineering.wordpress.com/2013/12/