TREXIN INSIGHT PAPER

May 13, 2015

A SLINGSHOT TO ADVANCED ANALYTICS USING BIG DATA

Achieving the Data Goal

The expectation around Hadoop and Big Data and the endless debate on investment and ROI has caught the industry's attention. This urgency has created more chaos in trying to deliver the promised data products that would solve business challenges. What is lost in this noise is a once-in-a-decade opportunity to get ahead in the game.

"The Game" here is: Intelligent data delivered at the hands of the business community to make early decisions and realize competitive advantage over peers.

This TIP focuses on outlining a roadmap to a Proof of Concept and also answers how to form an Advanced Analytics team and get the team on the ground and iteratively improve on the logistics in play.

For starters, as a practice at a minimum we need the following three players who can seed the foundation of an Advance Analytics organization using Big Data. For lack of better terms, we will call this team the "A Team" for the remainder of this publication.

1. BUSINESS SUBJECT MATTER EXPERT (SME)

Unlike traditional Business Intelligence which has teams of IT, Functional Business and Operations coordinating to get to a set of reports or a dashboard, Advanced Analytics relies on a unique set of definitive questions. The ability to ask the right question in the context of a business paradigm will win half the battle. This Business SME would have years of experience in the functional domain in question and would be considered a "Spotter of Opportunities" within an organization.

2. DATA SCIENTIST

This position is the key to connect the dots between the Business SME and the IT infrastructure team in order to predict the future. In plain English, this is someone who speaks about the evolution and maturity of the data from past to present, applies the math and statistical algorithms, and answers the question posed by the Business SME with a probable certainty.

There has been some amount of debate on what the skills and credentials of this position should be; many expect this role to be played by a PhD in Statistics or Economics. A PhD in the team is always good as a Data Scientist if the project is purely for mathematical or scientific research. However, someone with a PhD in Statistics or Economics may not be as willing to get their hands dirty with the IT world or could be too deep in the data woods for normal people to comprehend. Hence this generalization in the media seems a bit exaggerated.

The most important skill a Data Scientist must have is "story telling" abilities. In other words, observing the movement of data in time and deriving a reproducible story out of it that includes past, present and future. The key word in the previous sentence is "reproducible".

Nonetheless, a Data Scientist generally has multiple qualifications which sometimes includes being a PhD and should be able to bring all the information required to build the predictive model under one platform. Below is a list of behavior and orientation one would expect in a Data Scientist:

- **Passion about data**: A natural curiosity and passion to observe data and derive a story out of it. Existing data analysts with many years in the same organizations often have this quality.
- **Business knowledge**: Domain knowledge to comprehend questions that need to be answered and the ability to ask intelligent questions.
- **Knowledge of statistics**: A foundational academic background that includes math and/or statistics and knowledge of traditional Linear Regression models. Many in-house analysts on the operation side of the business have these skills; all that is often required is an online refresher course.
- **Hacking skills**: One of the key tasks Data Scientists do is data acquisition, which can come from multiple places within the organization or outside it. The Data Scientist should know how to collaborate with the IT infrastructure team to facilitate acquisition and movement of data in an available environment. In essence, Data Scientists need awareness and creativity to determine when and where to find the right data of the right quality.
- **IT knowledge**: To derive a predictive model and execute it on existing data. The key tools a Data Scientist uses are programming languages that enable this activity. Traditionally many educational institutions and organizations have used SAS or R as their enablers to do predictive modeling. Other languages that have gained traction over the recent years are Big R, Mahout, Python and last but not the least Apache Spark.

Needless to say this position is hard to fill; it is observed that it may be easier to train existing Data Analysts or Business Intelligence Architects and help them evolve to become a Data Scientist than to hire new ones from the market place.

3. IT INFRASTRUCTURE EXPERT

This position requires knowledge of MAA (maximum availability architecture) logistics. Invariably this is going to be someone who understands the Hadoop Distributed File System (HDFS). Why Hadoop? Because it is the only known framework that guarantees maximum availability for structured, semi-structured and unstructured data within the same logical place.

If the organization does not have an in-house expert in Hadoop, then training their front-line Application Architects is the second best approach. A good Application Architect can often come up to speed with the basic foundational skills in four to six weeks.

INITIATING THE FIRST PREDICTIVE/PRESCRIPTIVE DATA PRODUCT:

It might come as a surprise to many that Predictive Science was one of the first advents in the field of information technology back in 1960 right after the first series of IBM computers were invented.

Over the decades the scientific research community and clinical community in Europe and US have been using a series of proven statistical methodologies to progress in their field of choosing using Predictive Data products.

One may wonder if all of this has been happening for so many decades then what all the fuss is. The fact is, we can do all of this in batch and real time with interphases like Spark or Mahout that connect with HDFS frameworks that houses both structured, semi-structured and unstructured data. These very same predictive models that were originally designed many decades ago by our predecessors (clinical and research institutes) are now more effective in predictions because we can now run these model against petabytes of data and countless variables housed in HDFS frameworks. The statistical term given by the scientific community to elaborate this theory is called the Law of Large Numbers (LLN) or Central Limit Theorem (CLT).

As we have established the players that could seed the Advance Analytics team, here is a sequence of steps to initiate the first Predictive or Prescriptive Data product in an organization.

1. Identify the Business Unit (and representative Business SME) that needs forecasting -This could be the Sales team within an organization or a Finance department which usually uses many "factors" for its financial controls.
2. Getting the infrastructure in place - There are many companies that are providing packaged software and readymade Virtual Machines that house the Apache HDFS. However the two leading commercial market vendors are Cloudera and HortonWorks. These two companies have greatly simplified HDFS infrastructure implementation.
3. Let the "A Team" play ball - Create the project charter with the top five forecast questions that need to be answered and assign the "A Team" to the task. The team would first position the data from the business unit in question in the HDFS and build a data product to answer those forecast questions.

   Repeated success by the "A Team" will expand awareness in the organization and requests for predictions will grow. And as the volume in the HDFS grows the team will also grow organically and evolve to more complex machine learning methodologies.

CONCLUSION

Though some of this might appear complex at first, Advance Analytics in combination with Big Data will give companies the insights that would help them compete in the market place proactively. Business leaders will soon realize that the ROI is exponentially greater than the initial cost. It is just a matter of time when Advance Analytics in combination with Big Data will replace traditional Business Intelligence.

REFERENCES

Horton Works (http://hortonworks.com/industry/)

Cloudera (http://www.cloudera.com/content/cloudera/en/home.html)

Coursera (https://www.coursera.org/)

This TIP was written by Ashwin Manepalli, who specializes in enterprise architecture, management and analytics. Ashwin welcomes comments and discussion on this topic and can be reached at ashwin.manepalli@trexin.com