Big Data Road Map to Success for Defense and Federal Knowledge Management with Action Items

By Laurent Weichberger, Big Data Bear
OmPoint Innovations
[Revised May 2018]

Document Versions

Version	Updated By	Date	Comments
Draft	Laurent Weichberger	Nov-23-2016	Initial draft.
0.1	Laurent	Nov-28-2016	Continued writing.
0.2	Laurent	Dec-2-2016	Continued writing with input from training evaluations.
0.3	Laurent	Dec-5-2016	Added feedback and added use cases from Use Case Discovery workshop.
1.0	Laurent	Dec-12-2016	Added action items to each section.
1.1	Laurent	Dec-13-2016	Edited based on feedback.
1.2	Laurent	Jan-31-2017	Added more use cases.
2.0	Laurent	May-8-2018	Added more robust Use Case Discovery. General revision and editing for DoD and Federal usage.

Big Data Road Map to Success for Defense and Federal Knowledge Management with Action Items:

1 Determining the need: Use case discovery:

As a rule, all Big Data work starts with the need. In IT terms this need is depicted most often as a "use case" for how an individual actor (or system) interacts with Big Data. If there is no need, the project naturally will almost always fail. For this reason, we start with the need and discover and then develop the use case. In the situation at your Organization, we recommend you allow us to help you with our seminal "Use Case Discovery Workshop," which we have now successfully conducted since May 2017 around the world. This will help determine your actual needs. A sample of some use cases we helped discover and implement with Apache Spark can be found in *Appendix A: Organizational Use Cases*.

1.1 Big Data Stakeholders at your organization: Who wins?

Big Data at your organization means there are individual stakeholders who will achieve a victory when this work succeeds. A stakeholder could even be an entire department, if that department pioneers a Big Data technology and is known as a leader. We can help you to clearly identify the stakeholders at your organization, and if needed we can help conduct a separate set of meetings with just your stakeholders to ensure that their needs are being met, on a regular basis.

1.1.1 Known stakeholders:

Make a list of the obvious stakeholders and find out immediately what they need. If you are not meeting their needs with the use cases you discovery and develop, you will not achieve a victory.

1.1.2 Strategy vs. Tactics:

There is a difference between Big Data strategy and tactics. Strategy is more about why and when are we going to achieve this at your organization. The tactics are about how it actually happens.

1.1.2.1 Strategy:

1.1.2.1.1 Why?: not just because we can.

One of the major reasons "why" any company is embracing Big Data is due to relevance and ROI. The ROI conversation is outside of the scope of this road map, especially since DoD and Federal have a different definition of ROI than business. Suffice it to say, numerous organizations are saving significant amounts of time, energy, and money with Big Data solutions. In fact, in some cases they are "doing things" which they could not do before, due to the limits of their legagy technology. The frest Big Data technology choices provide hitherto unrealized benefits. The "relevancy" issue is extremely important, in that those organizations which fail to embrace and win with Big Data may in fact find themselves increasingly irrelevant. There are many organizations whose success seemed assured, and who are now found to be simply gone from the IT landscape.

1.1.2.1.2 When?: timing is everything:

The danger with many Big Data projects is that the timing is not right. By not right we mean either too early in the process, or too late. Too early may be characterized by problems with adoption, either with lack of support from the stakeholders and key players, or a lack of proper preparation by developers and users. There are many factors that contribute to proper timing. Too late could mean that the benefits and gains of the Big Data initiative are simply not going to make the difference and the window is closing on that Big Data project at the organization. The time for your organization to

embrace Big Data is now, and we are here to help you to achieve the initial small victories, and ultimately big wins. It is a delicate process.

1.1.2.2 Tactics:

Tactics means how it happens. In Big Data we need the right people in the right places doing the right things with accurate data.

1.1.2.2.1 Who and where?: which departments and people.

We are here to help you create a customized road map for your organization. This means, at least in part: Selecting the right people, with the right skill sets for Big Data work. This is not a minor task. In many cases, teams need a good deal of preliminary training and mentoring to ensure success. Assuming that your team will self-organize and teach themselves the relevant technologies, is most likely wishful thinking.

1.1.2.2.2 What?: Data sets available:

In any project, even with the right people, if the data sets are not available, or if some data is available but does not contain the most relevant data needed to solve the use case, the project will be in jeopardy. Therefore we stress "Use Case Driven Big Data," and also why the ingestion and data preparation phases are so vital to success.

1.1.2.2.2.1 Do we need different data?

To this end we need to analyze, and then ask, "Do we need to immediately start collecting data which is not present in our data sets today, so that we can support future use cases?" If the answer is yes, then this must be done carefully as a separate project, which will feed into future use case development. Again, this type of discovery is best done through the comprehensive focus on initial use cases, and prioritized future use cases. Preparing for the future will help guarantee success. Never assume the right data will be made available. Take responsibility to ensure it is available.

1.1.2.2.3 How?: Tool selections (technologies)

Observe carefully the current and proposed future tool selections around this Big Data work. Ask yourself if these are indeed the best choices, and let us help you with another perspective. There are many Big Data tools available, and it would be negligent not to review and contemplate for each Big Data project which tools best fit that Use Case (problem) and the proposed solution. (e.g. not all problems are best solved with Apache Hadoop). Also, many of the Big Data technologies have an overlap of functionalities. Some are more current, and vital. Some are already legacy and dying out. We will share more about tools in a later section of this road map.

1.2 Leadership: Establishing an internal "Big Data Champion":

In addition to having a Big Data team itself, we feel strongly that having Big Data leadership at your organization, as reflected by the presence of one or more "Big Data Champions" is a strong move. Such an individual should be given this opportunity in addition to their normal "day job." If possible, this person should come from within existing teams, and not hired from outside. While it is okay to create a new hire "req" for this – it sends a different message, one of – we need outside help. The culture we seek to create is one of internal empowerment, and investing in your existing resources to bring them all up. We believe this for the following reasons:

1.2.1 Shows immediate support for Big Data initiatives from the Stakeholders.

1.2.2 Immediate allows your organization to invest in individual leaders with extra training, mentoring and career enhancing work, which in turn increases morale.

1.2.3 When done properly, that Champion will naturally evolve into a Big Data mentor for those who join the Big Data team later and naturally have less experience.

1.3 Psychology: "In the Shadow of Big Data":

In our experience, there is a psychological aspect to all Big Data work. When we speak about the "shadow" in Jungian psychology, we generally mean that which is deemed unacceptable, or unworthy, or unspeakable. In the case of Big Data at your organization this may mean such things as fears around Big Data manifesting in different employees or even contractors. Such fears are natural, but to ignore them would be a disservice to the teams and to your organization. Examples of such fears will be given here next. Please keep in mind, we can help alleviate these fears with proper Big Data education and mentoring:

1.3.1 Fear of appearing to be ignorant (not knowing) what one needs to know for success in one's given role.

1.3.2 Fear of losing one's job.

1.3.3 Fear of becoming irrelevant in one's chosen field, if one doesn't immediately retool.

1.3.4 Fear of earning less, or getting a pay cut, or not getting a desired bonus.

1.3.5 Fear of being overwhelmed (and lost), or non-functional, in this technology space.

1.3.6 Fear of not knowing what one doesn't know – not even knowing what questions to ask or where to go for help, or how to even start.

1.4 Action Items for Section 1:

1.4.1 Establish Organization Big Data Champion Role, to add weight to this program.

1.4.2 Create Seminar: In the Shadow of Big Data, to immediately help alleviate perceived psychological tension.

2 Technology Selections: "Beyond Hadoop":

It is quite common in the IT industry for people to believe that Big Data equals Hadoop. While Apache Hadoop is an important software tool in the Big Data ecosystem, it is simply one of dozens of tools available. Without a proper understanding of this ecosystem, and the various types of Big Data problems (use cases) and their solutions, there is a danger of thinking that Hadoop will solve all Big Data problems. Hadoop was never designed to be all things to all people, it was designed for a specific purpose (known as HDFS and Map/Reduce), and it has evolved from there.

2.1 Proof of Concept: Some technology.

It will be clear that your IT team was responsible for creation of an organizational Big Data proof of concept. That is wonderful, yet we need to review the technologies choices, and the proper use of that technology. For example, I was recently at a large bank (which shall remain nameless). They were improperly using the Apache Hadoop HDFS technology, causing frequent crashing of the NameNode server. The root cause is now an "Anti-pattern" or what not to do. They asked me to help them fix this problem. Better would be that at the outset they didn't misuse the technology. I wish I

had been there at the start of the project to guide them, however they called me when they were in pain and suffering. Knowing the anit-patterns is important. We can help guide you.

2.1.1 Should your organization consider other technologies?: At some point, when you are ready, it would be wise to learn about the other technologies available in this space, and see how they are currently being used by large companies -- such as Nike, eBay, PayPal, Netflix, Apple, and others. Regardless of the government aspects of the work involved, we can still learn from the experience of those who came before. What follows is a very brief introduction to just a few of the major players in the Big Data space:

2.2 Column Family: HBase and Apache Cassandra:

HBase and Cassandra are known as “column family” databases.[4] They are very fast and efficient at storing Big Data, and allow companies (such as Netflix) to store all their metadata for their entire customer base of over 50 million users. These are both open source technologies which use a NoSQL key-value data store as the core technology. DataStax is the vendor that offers Cassandra as a product, and the vendor MapR has done the most to support and build upon HBase.

2.3 Document Oriented: MongoDB and Couchbase:

Many companies want to process Big Data in as-simple-a-way as possible, and for certain projects a “document” database is very useful. When we say “document” we typically mean storing data as JSON text documents, while other document types are allowed. MongoDB and Couchbase are two vendors who have mastered this space. Nike revealed to me personally that they selected Couchbase as a solution for one of their projects because of “ease of use.” I share this story because they have some of the finest engineers I have met in the field. Use what is best for the job at hand.

2.4 Apache Spark: A better mousetrap:

No Big Data conversation today can be complete without mentioning Apache Spark. I am one of the only Databricks Certfieid Apache Spark instructors in the world, and I have been teaching globally since my first teach at Booz Allen Hamilton (BAH: 2015 & 2017). I also helped BAH prepare for their Spark Certification, and wrote seminal “Spark cert prep” materials for them. I also taught publicly at Strata+Hadoop World (2015). In addition, I have taught over twenty on-site private Apache Spark classes for customers including but not limited to:

- USAA
- RAND Corp.
- IBM
- Royal Bank of Canada (RBC)
- Tesco (UK)
- Hastings Direct (UK)
- Llyods Bank (UK)
- GM Financial
- Canadian Pension (CPPIB)
- CVS
- Pekin Insurance
- Mutual of Omaha
- Charter/Spectrum
- Carestream

2.5 Graph Databases: Neo4j, et al.

Some use cases are best solved by using a Graph database. The graph technology is significantly different from either a Relational Database or Hadoop. In Graphs there are data models of relationships using Vertex (Node), Edge (Connection), and properties in a network of links. This is a growing field, and Neo4j is a leader in this space.[5]

2.6 More Layers of the Stack: Datameer, and others

At one organization, we have already taught the teams how to export data from Datameer to Tableau for example. The more one works in this space, the more it appears to be a tapestry of various technologies that are used to get one's work done. To narrowly focus on any technology (including Hadoop) in the long term at your organization is simply not wise. We should become more and more aware of the available technologies and continuously improve by reevaluation of technology choices and available technology stacks. All of this is continually improving, evolving, and new technologies are emerging frequently. Embrace this change.

2.7 Action Items for Section Two:

2.7.1 Create seminar on Big Data technology stacks beyond Hadoop.

3 Roles and responsibilities: who does what?

3.1 Your organizational staff:

For the purposes of this road map, we are going to distinguish between organizational staff, contractors, me at OmPoint, and your many vendors. This will help us sort through the various issues required to achieve Big Data success.

3.1.1 Your Team:

The primary understanding about the needs of the your team must be communicated to OmPoint, so that we can help you to create star players in this Big Data field. We want to encourage and support all of your efforts, as follows:

3.1.1.1 Big Data Rockstar:

Insofar as possible we should help the your team, and then all other teams who wish to follow, to create what could be called a "Big Data Rockstar." Once there is one such star, there can easily be more, but if we focus on investing in and shepherding one individual to such success it will inspire others. Let's us make some recommendations here:

3.1.1.2 Datameer

I have experience with this tool, since I worked there. Thorough training in Datameer including as much of the technical training as possible (based on the background of the individual) would benefit organizations who have few or no "development" staff. Datameer accomplishes much through the use of their UI so that you don't need to write raw code. Such Datameer training from us would include but not be limited to the topics: *Custom Function Creation, REST API, and EventBus Listener.*

3.1.1.3 Hadoop

Since Datameer completely relies upon Hadoop as the engine (and in Datameer 6.x Apache Spark as well), we feel strongly that Hadoop training is important for Datameer users. Such training would simply raise the understanding of the Rockstar to the level needed, without requiring them to be "too technical."

3.1.1.4 Other tools:

As previously mentioned, there are many Big Data tools available in the ecosystem, and discovering which tools are best for new use cases is part of this journey. This is where Big Data mentoring comes into play. We can help mentor such a star as they learn and apply other tools, beyond Datameer and Hadoop.

3.1.2 IT Team: Continuous Learning:

In the case of the IT teams at your organization, OmPoint will need to be informed about ongoing learning needs. Such a meeting and discussion would only benefit the IT team, and we can review this on a regular basis. Perhaps IT feels they know, "all they need to know" for the time being, or perhaps there is a pressing need of which we are not aware. Let's all review the needs on a regular basis: monthly at first, and then on a quarterly basis as we get to steady state for your org.

3.1.3 Big Data Champion:

Separate from a Big Data Rockstar, we envision that the Big Data Champion role be available within the organization as a role (in addition to the employee's normal job description) whereby they take on a persona of leading and shepherding the organization in this direction. Think of this role as a Big Data change agent, to be given extra-special attention in the way of education and mentoring from OmPoint. Ultimately, this Champion, and others who embrace the Champion role will be the mentors to those who come later into the Big Data Team. We feel that communicating this as a role to be conferred upon a full time staffer will help morale around Big Data and send the right positive message.

3.2 Your Contractors:

For the purposes of understanding Big Data success at Organization, can we ask some vital questions and make some technical assumptions?

3.2.1 What is the difference?

Should organizational contractors be treated any differently than Full Time Employees (FTEyou're your organization?

3.2.1.1 Access/Security

We know some datasets contain sensitive information about individuals. Furthermore some data is deemed sensitive because of the possibility of "insider leaking" of information. Some data is limited to a certain set of people who have permission to see it ("need to know"). Sometime we only show a sub-set of the data during our educational sessions. To this end we advise a Big Data policy and procedure around data access and security which clearly delineates these and other related issues when it comes to how an organization's contractor gets involved with Big Data.

3.2.1.2 Responsibility

Can we safely say that the responsibilities of an organizational contractor are the same as the equivalent role of an organizational FTE, regarding Big Data, or do we need to spell out the differences in the policy and procedures? This is something to consider so that should a contractor be moved off the project, or terminated, they are not leaving with some proprietary information that is not covered in your normal policies. We have precedence for these issues in news, and I have my own personal understanding of these issues, which I should not write here, but which I can explain in a meeting.

3.3 OmPoint:

OmPoint Innovations is dedicated to your organization arriving at Big Data success. This road map from OmPoint aims to fully empower your organization's teams to achieve repeatable victories. This is proved by the fact that we are freely sharing this information with you at the DoD and Federal Knowledge Management Symposium. You can do what you wish with this information.

3.3.1 Make the best of what we can offer you:

In order to achieve the wins, we envision, we feel that your organization should make the best of what OmPoint has to offer in the way of Big Data training and mentoring, especially during your "transition time," when the psychology of the small victories will be most important.

3.3.2 Goal is to become independent:

Believe it or not, ultimately the goal from OmPoint's point of view is that your organization should become fully empowered and independent in the Big Data space. You will thereby no longer need to rely upon OmPoint. This may be a poor business plan, but it is the healthiest for all. There are many steps and stages to this Big Data process, and this road map aims to clearly define them all.

3.4 Big Data Vendors: Datameer, Databricks, DataStax, etc. (notice a pattern here)?

3.4.1 Once you go down a path with a particular vendor, you enter into a profound relationship. Why do I say that? Because I have worked at these data companies, and I know their cultures, their pros and cons, their ways and means. They are not charitable, and rarely "care" about government or defense initiatives. They are simply in it for the money. Selecting the right relationship is important, because you have to live with them. Live with here means support, documentation, examples, professional services, etc. They are not all created equal (repeat NOT).

3.5 Avoiding the doubles-tennis problem.

One of the biggest problems in any project of this nature is what we call the "doubles-tennis" problem. This is a classic problem wherein two players on the same team assume that a ball which is hit towards them -- *but between them* -- will be handled by the other player. This is primarily due to non-communication. The result is that neither player makes any motion towards the ball, and instead they both watch the ball travel between them, while the opposing team scores a point. In the case of Big Data this is slightly different in that lack of proper communication means that ball is an "issue" which, due to false assumption, is not handled and slips through the cracks. This is easily solved by practice and ownership communication, such as, "That is yours!" or "I have this one." **The Agile-Scrum process for example is great at assigning ownership to issues.** If both players know how to communicate and move forward in ownership, it works out great. In the end, it is far better for both players to try to solve for it, than to assume the other will work on it. What is best is that one takes ownership with support from the other. We can solve this with a clear communications process. Laurent is also a Scrum Product Owner, and ScrumMaster and can teach Agile-Scrum to your teams.

3.6 New Hires: when there is headcount.

I have often been told by clients, "There is a Big Data Team forming soon…" and then the questions start coming about Big Data new hires. The typical answer is, when there is "head count," we can hire. The best outside resources to hire have a Data Science background, or they already know Hadoop and Spark, or whatever is already in place as your organization's existing technology stack. Often, however, there is no clear Big Data leadership (as previously outlined) and the team doesn't materialize or fails continually. We can help you here.

3.6.1 Data Scientist: Machine Learning and advanced analytics:

Finding a Data Scientist in the USA at this time is not so easy. Organizations around the world are slowly waking up to Big Data, and whenever possible pulling in resources. The ones already awake

have gobbled up the best and brightest. The time to start the search for a Data Scientist or Machine Learning specialist to perform Big Data advanced analytics is right now. Don't delay.

3.6.2 Other Reqs?

Beyond a Data Scientist, are there other reqs that should be opened? That is something we need to consider carefully when creating the Big Data team. We have considered some of this already, however we suggest that having a thorough meeting about your new Big Data Team would benefit your rganization. Allow OmPoint to offer guidance on this point.

3.7 Action Items for Section Three:

3.7.1 Have a "Town Hall" style meeting whereby all Big Data players are invited: Your Org Staff, Contractors, OmPoint, other vendors (Hortonworks, etc.) and share the road map, the roles and responsibilities, and open it up for questions and answers from all. This will help foster the transparency and trust, both of which will be needed for success.

4 Adoption: top down and bottom up (meet in the middle)

4.1 Executive level sponsorship and support for Big Data.

Without the right levels of sponsorship for Big Data within the organization, the Big Data projects will get bogged down in politics and infighting. It is up to the org sponsors to clear the way for the success of Big Data.

4.1.1 Leadership support: right messaging

Insofar as possible it is wise for the highest ranks of management to make statements (internally and externally) about the Big Data initiatives at your organization. This will boost morale and send the positive message to all who need to hear it.

4.1.2 Manager level issues: right vision

At the EVP, VP and Manager levels (basically all the stakeholders) we need to see everyone singing from the same sheet music for this to be a success. We feel that having a regular meeting with the Big Data stakeholders to address concerns, clear up any issues, and make sure needs are being met, is extremely important especially at the early adoption phase of the work.

4.2 User level adoption: right usage

Of course, at the end of the day, user adoption is the key to success. Even if all the stakeholders love the Big Data initiatives, they will ultimately rely on the users of the software tools. We can help achieve right usage by the users through continued education, mentoring and creation of Big Data Champions.

4.2.1 Hadoop is not intended to replace SQL.

One way that we can allay any fears that crop up around usage is to clearly state that Hadoop (and other Big Data technologies) were never created to "replace SQL." This has come up in every training session at I have taught. Part of the psychological work we referred to earlier is to give the proper perspective about where we were, where we are, and where we are headed. I can assure you SQL is not going to vanish anytime soon. People have their customary way of working, and getting results. We don't want to threaten that, or suggest that we are going to take away any of their tools. We are here to add to their tool set, and further invest in, and empower, them to get their work done faster, and more efficiently -- that's all.

4.2.2 Healthy outlook

A healthy user outlook incorporates all the above and more. It is mature in seeing how old technologies (like relational databases), and tools like SQL queries and SAS will remain, while newer technologies rise up and in some cases, will eclipse the old ways of doing things. There is nothing to fear, and yet we want to allow everyone their natural process of adopting Big Data in their own way, and at their own pace.

4.3 Action Items for Section Four:

4.3.1.1 Communicate with the different layers of your organization's employees, from the highest levels down to the users, to confirm support for the Big Data road map. Then plan and execute this road map. Listen for concerns and capture feedback from these different levels of experience. Incorporate feedback in the road map. Rinse and repeat.

5 Implementation: "Never enough time to do it right, always enough time to do it again."

5.1 Installation:

Big Data installation is often complicated, and if not done properly can lead to chronic problems. In addition, once the installation is successful, almost all technology stacks benefit from performance tuning. This tuning can be an arcane art. Don't attempt this alone at home, get some help.

5.1.1 On premises vs. Cloud (AWS, etc.):

One of the first considerations that must be resolved is whether to build out the technology on premises, or in "The Cloud" (such as Amazon Web Services). Of course, your policy and procedure may limit this discussion. As time goes by, it would be wise to review this decision. There are trade-offs, between on premises and cloud solutions, and those should be carefully reviewed.

5.1.1.1 Selecting the right node type (memory, CPUs, etc.)

Only after deciding the location of the technology, can we determine which node type (hardware requirements) are best for this particular technology. For example, an Apache Cassandra cluster node might benefit from a different node type than an Apache Hadoop node, and etc. In any case, selecting the node type can mean the difference between slow jobs, fast jobs, and flat out job failure due to exceptions. Fatal jobs have been seen in the Datameer product when the node type is too small to meet the requirements of the work. Once the node type has been selected, this too should be reviewed on a regular basis.

5.1.1.2 Selecting the right number of nodes

In the same way that node type is important, the actual number of nodes which make up a cluster for a specific technology stack is vitally important. Once the "replication factor" is decided, and the various storage, CPUs, and other issues are decided, the initial number of nodes and the plan to increase the number of nodes over time should be decided, or at least planned. These plans can change; however, we can predict how much data will be ingested into your cluster over the next year, two years, five years, and etc.

5.1.1.3 Security considerations.

As previously stated, it is important to keep security considerations front and center especially in the DoD and Federal sectors. The easiest way to be hacked is to ignore the need for security. To this end OmPoint can perform a "Security Gap Analysis Report," based on what has already been done by your organization, and what else should be done.

5.1.1.4 Performance tuning:

Performance tuning means making an existing functional technology stack more performant by tuning the various variables available in the configuration settings for the layers in that stack. Some examples will be given below; however, they are not intended to be an exhaustive list. The point is that once the system is working, a concerted effort to refine and tune the systems should be engaged. This process is known as performance tuning, and is specific to each layer as follows:

5.1.1.4.1 Hadoop[6]

Hadoop has many configuration parameters, which are well documented, and which if set properly will greatly affect performance. Whether it is about RAM and buffers, or which directories are being used for effective storage there are a lot of considerations, and this should be a separate project that is regularly undertaken.

5.1.1.4.2 YARN[7]

Tuning YARN is a separate set of considerations that focus more on the YARN Container configuration settings. Cloudera provides a fine YARN tuning spreadsheet.[8]

5.1.1.4.3 Datameer

In versions of Datameer prior to 6.0, Datameer relies upon Apache Tez. Datameer documentation speaks to performance tuning for Tez and other configuration parameters which affect Datameer performance.[9] In the same way that we need to have a Hadoop/YARN performance tuning project, the Datameer performance should become a project.

5.1.1.4.4 Spark (in Datameer 6.x)

In versions of Datameer after version 6.0 there will be greater reliance and dependency upon Apache Spark. The configuration parameters for Spark are *different* from the Tez settings. This should be taught and understood by any organization. Laurent used to teach Spark internally to the Datameer staff, and therefore he remains in a perfect position to teach the same understanding to the any organization's staff.

5.2 Training:

At OmPoint we believe in training and mentoring at various stages of the road map. We have already performed dozens of training sessions on-site globally, and we are available to continue offering training sessions as and when deemed best by your organization.

5.2.1 Selecting the right team members: capabilities

Training the right people is always an important consideration. Some students who are selected may feel they are not suited for training, for example they may feel they lack the proper background to succeed with Big Data. In such cases, we can either create a course to help prepare those who need more preparation, or we can review carefully (beforehand) the student aptitudes. Pre-training assessments can be created, etc.

5.2.2 Big Data Training: Thinking in Big Data

A big part of learning Big Data is learning to think differently. Believe it or not, students who have little (or no) data background seem to learn Big Data technologies much more rapidly than those who have an extensive relational (SQL) data background. Why? There is much less to un-learn. We are now seeing students come out of university with zero relational data understanding. They have only ever worked with NoSQL and other Big Data technologies. When shown relational tables, and

Primark Key/Foreign Key relationships they ask, “Why would you want to do that?” It is a fact. So, the goal of many of the training sessions, at least in part, is to help the students to think in Big Data.

5.2.3 Basic Orientation:

We suggest that a basic orientation at your organization around Big Data consist in at least one full day of training in the technologies giving a Big Data background and some exposure to the actual toolset. To this end OmPoint has already performed such training successfully many times.

5.2.3.1 Minimum one day: how to use the tool(s)

The one day beginner agenda includes the rationale of Big Data, the various approaches (at a very high level), introducing Hadoop and a few other concepts, and then diving into a practical session on using a tool of your choice to solve real use cases.

5.2.4 Power Users:

The Power User training is more advanced and is focused on how best to leverage specifically advanced functionality of your chosen technology, with various developer APIs.

5.2.4.1 For example, Datameer REST API: already in Datameer 5.x

There is a robust REST API which allows users to request Datameer to perform tasks from outside of Datameer, via REST calls. This is Datameer REST API has already been tested by OmPoint.[10]

5.2.4.2 Custom Functions: already in Datameer 5.x

When the out of the box functionality from Datameer is not enough, there is a “custom function” API which allows you to write your own reusable functions. OmPoint successfully taught how to write custom functions during the Datameer training in Dallas (November 2016). This courseware is already written and can be reused any time.

5.2.4.3 EventBus Listener: already in Datameer 6.x

For those who require more advanced “Auditing,” Datameer provides an EventBus Listener API in versions newer than 6.0. OmPoint can teach the EventBus Listener API at any time. This material has already been written.

5.2.4.4 Search based: ElasticSearch-Hadoop integration

The question arose regarding whether Hadoop supports “search based” use cases. After some research the answer is, yes. There is a project to integrate Datameer and Hadoop with search via Solr, and Elasticsearch-Hadoop.[12]

5.2.5 Learn about other tools, technology stacks.

While the above training sessions are ready to offer your organization, there are other technologies available to your org as well. Preliminary training in any Big Data technology is not only possible but wise. Knowing all the technology possibilities can only benefit those making difficult Big Data decisions. Otherwise, as the saying goes, “If all you have is a hammer, everything starts to look like a nail.”

5.3 Support system:

As the Big Data work at your organization matures, there will be a need for a predictable and stable support system.

5.3.1 Internal support levels: IT, OmPoint

Level 1: Your organization’ internal support. We can help them revise their understanding.

Level 2: OmPoint support from Laurent Weichberger, et al.

5.3.2 External: vendor support (depending on your vendors)

Level 3: For example, Datameer technical support. This includes opening tickets at support@datameer.com, etc.

5.4 Use case driven development:

We have created a seminal Use Case Discovery process, as follows, and we share this freely in hopes that you will allow us to help you implement this in your ongoing projects:

- **Use Case**: (noun) "A description in simple language regarding the nature of a journey through software, including but not limited to what the user of the system experiences, what the system responds with, and any other details but remaining higher level than fine grained task interactions."
- **Discover**: (verb) "To make known, to reveal. To uncover and shed light upon in order to better understand the nature of a thing."
- **Agile-Scrum** (noun): "A successful agile software development methodology, practiced worldwide, using simple constructs and roles to maximize efficiency."

At his first Spark teaching for Hortonworks, Laurent offered his seminal "Use Case Discovery" workshop. The goal was to tackle a real Big Data business use case during the last day of training. This has proven to be wildly successful May 2017, as follows:

- We ask seven **discovery** questions...
- We use Agile-Scrum.
- We nominate a Scrum Product Owner
- Together as a team we write "User Stories" and prioritize them.
- We create thirty minute "Sprints" (during which time team members are always working).
- We apply any technology: Apache Spark, Hive, etc. to solve the problems.
- **Tremendous success! All in one day of work, some examples of our success:**
    - Cox Cable: **Saved $50K with one day workshop.**
    - Charter Spectrum: **Increased analytics speed going from raw python to PySpark:**
        - **From 5 hours** to ***12 minutes*** *(that is a 25 x increase).*
    - Mutual of Omaha: Implemented a **GeoLocation Engine** for production usage.

5.4.1 Step One: Seven Seminal Questions:

1. Short Name? (means, if I am asked, what do I say I am working on when we are on the elevator, and about to get off at my floor):
2. Problem Statement? (long description): That which we trying to accomplish or solve. Don't include a solution, just the problem please. Harder than you think to focus on just the problem, and get agreement from team.
3. Proposed Solution?: How do you propose we solve this? Remember beginner's mind. Leave your legacy thinking at the door. Any technology, and any solution is permissible.

4. Data Sources?: You must identify your data sources to succeed.
5. Stakeholders?: Who cares about this project enough to want it to succeed?
6. Reviewers?: Who will take the time to actually review your development work and say, “Great!” or, “This won’t work because, xyx…”?
7. Further Thoughts?: Any juicy ideas that you have which don’t fit nicely in one of the above categories, but may be worth keeping?

5.4.2 Agile:

- With Agile we have roles, responsibilities, artifacts and process to adhere to, like this:
- Roles:
    - **Product Owner**: Author of User Stories (based on the use cases), generally not a team member that develops software but we can hybridize that role at CPPIB.
    - **ScrumMaster**: facilitates the process and the working of CPPIB team with Product Owner (every Sprint).
    - **Team Member**: Implements the stories as working runnable software.
- Process:
    - Creates Sprints with Team (e.g. a repeating cycle of work lasting until all work is “done”). Typical Sprints are two weeks in duration, normally it will be **30 minutes each.**
    - Creates “definition of done” with team.
    - Works with Product Owner (one or more of you) to define the **User Stories**.
    - Removes obstacles (like Ganesh) from the path of any team member with a block.
    - Artifacts in Scrum include: Product Backlog (all user stories), Sprint backlog (what is being worked on currently), burn down chart (if story points are involved), etc.

5.4.3 Step Three: User Story Creation:

- A user story is a document which can be “pulled” by a Scrum Team Member to be implemented as one or more tasks.
    - A Team Member can *never* be assigned work, ever.
- User story creation is the responsibility of the Product Owner(s), as they know the most about the problem domain and the desired solution.
    - A story can be an “Epic” (must be broken into sub-stories).
    - A story can be just a story (small enough to be pulled by a Team Member).
    - A story *cannot* contain task level detail (see below):
- Team Members know the most about the actual implementation of the solution:

- Product Owner *may not write task level items* but must TRUST the Team Member who pulls the story to decide all tasks needed for proper implementation (*no fingers in the team pudding please*).

5.4.4 Step Four: Implementation - Putting it All Together

1. Self organize around answering all of the seven discovery questions.
2. Allow ScrumMaster (Laurent in this case) to help refine the answers to the questions until the entire team and ScrumMaster believe it is ultra clear and true.
3. Self organize around a Product Owner: ”Oksanna,” knows the most about this problem domain and desired solution.
4. As a team, focus on the proposed solution:
    1. Write as many stories as necessary to implement the solution.
    2. Provide priorities for each one of the stories (1= highest priority, 5=lowest)
    3. Self organize with guidance from ScrumMaster:
        1. Each team member pulling a story (more than one team member can be working on the same story, in which case the team members solve that story together).
        2. Team Members must create a task list (one or more tasks for that story)
        3. Implement the solution
5. Self organize around integration and testing of all of the above.

Real World Success Story:

- Before the use case discovery workshop the PySpark code took *30 minutes* to run. ☹
- After use case discovery refactoring the PySpark UDFs to use the pyspark.sql.functions, the PySpark application version 2.0 took **11 minutes** to run. ☺
- **Success! Goal was 50% increase in speed… we achieved 63% increase. Yay!**

5.5 Use Cases: Context

In the context of Big Data at your organization, use case driven development means that we know (or discover) the data use cases before we perform the data ingestion. The reasons for this are many, but not the least of which are due to the usefulness of “schema on read” whereby the column name, data type and inclusion or exclusion of a column are determined when ingesting data. Without the use case guiding this effort, there is a tendency to “ingest all” data, which makes it more complicated downstream when trying to make sense of this vastly ingested data.

5.5.1 Importance of use case discovery

The importance of use case discovery lies in the need for your organization to work effectively and in an agile manner on the prioritized use cases, which will be most impactful. Without use cases, there is a tendency in the organization to resort to "ad hoc" only requests, or be very reactive (management by crisis). With prioritized use cases, there is a responsive culture, and the culture that emerges is more mature and work is made more enjoyable for the teams. In addition, use cases can be tied to specific metrics, and ROI, and one can determine (over time) the value added by certain projects over others. This will become the Big Data legacy of your organization. We are aiming for victories in Big Data, and tying the victories to use cases will be a vital aspect of this new culture.

5.5.2 Organizational Inculcation

OmPoint has already experienced some resistance to use case discovery at many organizations. This is natural, especially when there has not been a legacy culture of use case driven development. With proper guidance, we believe that any organization can and will embrace this use case culture, and will benefit tremendously therefrom. OmPoint is committed to holding regular "Use Case Discovery" workshops until this is second nature for all teams. Our goal is for OmPoint to be no longer required to initiate and lead in this area. To this end, some use cases generated from our Use Case Discovery workshops are found in Appendix A of this document. For a comprehensive set of use cases, read my blog at: https://www.linkedin.com/in/bigdatabear/detail/recent-activity/posts/

5.6 Ingestion and Data Prep:

Before data analysis can be performed, there is an ingestion and data preparation phase. This should be considered carefully, as the culture around these steps has a way of perpetuating itself, in a sort of "That's the way we've always done it…" kind of way. The rules and culture established early on in this Big Data process tend to have a life of their own.

5.6.1 Ingest All vs. Schema on Read

As has been previously stated, we have seen (and discouraged) a culture of "ingest all" at some organizations, whereby the schema on read functionality is rarely used, or it is used in an ad-hoc way based more on the personality involved in a project, rather than based on some rules or policy and procedure. This is somewhat dangerous. For example, one person creating a Job in Datameer may decide that it is necessary to change the column names when ingesting, from whatever the original data source names were to something that they find more clearly understandable. This is fine in one way, but it opens up the door for each person involved to have their own naming practices (e.g. one person uses *camelCaseNaming* and another uses *underscore_separator_characters* in naming). Also, if the schema on read functionality is ignored, then all columns associated with a data source are ingested even if NOT ALL THE COLUMNS ARE NEEDED for a use case. Again, these are trade-offs. Again, for example when performing a Datameer ImportJob the data is stored in Hadoop's HDFS, so the footprint associated with each ImportJob will be substantially different depending upon what is decided here. With use case driven development, we recommend not ingesting all data, but creating all work based the requirements in the use case. This leads to clarity and rapid solving of use cases.

5.6.2 PII considerations

No data road map would be complete without addressing the issues around "personally identifiable information" (PII). We need to share with all Big Data teams the policy and procedure around PII. It has come up in training that the datasets have to be created as sub-sets to ensure no PII or security community rules are being broken.

5.6.2.1 Data masking.

It is possible to "mask" PII data with some replacement scheme, whereby all characters are transformed or obfuscated into "*" or the first part of the data is in some way transformed to protect the data. While these practices are easy to create, it is important to establish rules and a culture that enforces usage. We have a blog about this practice using Apache Spark to obfuscate PII.

5.6.3 Partial Datasets (range bounded)

Datameer allows for dates to be used to create range bounded datasets, known as Partitioning. This is extremely helpful, however this must be done at either the time of ingesting the data into Datameer or when saving the data to HDFS. We find many reasons for partitioning the data (such as faster time to POC, and solving use cases on smaller datasets before using the entire data set). We would like to advise that ingestion be partitioned when possible. (Non-date related partitions were not yet available in Datameer at the time of this writing).

5.6.4 Ingesting external data:

Some use cases rely upon the ingestion of "external data" into the tools. By external we mean here that the data was neither created, nor is it owned by, your organization. Weather data is a good example of this, which may be used to solve a use case asking about how weather affects your organization's use case.

5.6.4.1 Policy and procedure

You should have a policy and procedure in place for how and when it is allowable to ingest external data, since data ingested with these tools does create a storage (e.g. HDFS) footprint. Do you really need your own copy of this data?

5.6.5 Ingestion Validation

During 2016 we encountered a severe problem related to ingestion and time zones. This was an unfortunate problem, however upon solving it we realized there was a lack of validation process in place. To this end we created a comprehensive Data Dashboard and Validation procedure to avoid having this issue recur. We demonstrated the effectiveness of our process, and can share that with you.

5.6.5.1 QA, Test Plan, UAT

The validation process (mentioned above) should be a part of the larger Quality Assurance, Test Plan and User Acceptance Testing process that we can help you put into place for your organization around the Big Data project work. We will create a separate document sharing the QA, Test, and UAT processes upon request.

5.7 Analytics: An Agile approach:

For those not familiar with Agile, there is a development movement which clearly states an agile methodology which aims at rapid results using an iterative approach. The "Scrum" methodology is an excellent Agile way of working, and OmPoint can help teams learn Scrum if needed.

5.7.1 Rapid results.

By rapid results we mean that the Agile-Scrum methodology can produce results in as little as a day or two, depending on the length of the "Sprint" (a cycle of development work). Each Sprint is an iteration of the work. Work can be revised and integrated into a larger project deliverable. We are big proponents of Agile-Scrum.

5.7.2 How to ensure no two people are working on the same problem.

Agile-Scrum is also a great way to ensure that no two people are accidentally (or unintentionally) working to solve the same problem. In highly "siloed" organizations, it is common for the communication to be *so poor* that two employees may be trying to solve the same problem, even writing similar or identical code, simply because they don't realize that there is a duplication of efforts. Agile-Scrum's "Daily Standup" meeting and "osmotic communication" aims to overcome these weaknesses.

5.7.2.1 Communication on who is doing what and when.

At the very least, your organization will benefit from any communication mechanism which clearly delineates who is doing what and when, to avoid duplication of efforts, and wasted work. Furthermore, use case driven Big Data work, which has solved use cases, can be the source of a Use Case catalogue, ensuring that the same problems are not re-solved by teams over time.

5.7.3 Has the solution been created already?: Don't reinvent the wheel:

To this end, any way of cataloging problems, solutions, and source code will help teams over time to be as efficient as possible. We touch up on this further in the Data Governance section of this road map. There must be a single source of truth in the way of documentation on what has been done, to enable efficient reuse of your organization's created solutions.

5.8 Clear communication: How to avoid politics

A wise person once said only three people are required for politics. Naturally, in an organization the size of your government organization we have (and will continue to experience) political situations around Big Data. This is natural and to be expected. Anyone who says that there are no such politics is either naïve or in denial. A mature attitude says, okay given the reality of politics in any organization, "What can we do to minimize the impact of these politics on the Big Data projects at hand?" To this end we have two important recommendations:

5.8.1 The use of Non-violent Communication (NVC), sometimes known as Compassionate Communication.

5.8.1.1 Non-violent (Compassionate) Communication: NVC.

The NVC communication process was initially developed by Dr. Marshall Rosenberg.[13] The simplicity of the process holds its power which has been proven worldwide since it was first published in 2003. The process says that the best communication be shared in four steps, in this order:

1. An observation.

2. A feeling about the observation.

3. A need related to the observation and feeling.

4. A request.

This may sound simple or obvious, however in practice this has served to clarify and disentangle many otherwise combative communications, which usually end up causing days or weeks of political turmoil. We would like to suggest that at the very least Organization consider some practice of clear communication around Big Data needs, if not NVC, then some other communications discipline. OmPoint is happy to create and offer training in this area for Organization teams.

5.8.2 The Anatomy of Peace

Another amazing tool in the area of communication and conflict resolution is the book titled: *The Anatomy of Peace*, by the Arbinger Institute.[14] This book outlines various ways in which people and organizations get it wrong in relation to how people communicate and treat each other, and offers real tools for treating each other better, proactively. This is a fantastic resource for anyone, especially teams.

5.9 Frequent small victories: the psychology of winning.

From the psychological standpoint, there is a great deal to be said about winning in Big Data. In this context, we mean that a team of users can achieve one or more small victories using Big Data technologies to solve real business use cases, and claim victory. Once this is done a few times, the morale of the teams around Big Data usage is boosted. Without these small victories early on, we have seen morale erode and doubts set in. To increase the likelihood of success we have found that creation of a Big Data user group, and learning lab, are extremely helpful.

5.9.1 Big Data User Group: where to go.

The purpose of the organizational user group will be to create a forum where big data users can go to share their questions, problems, share success stories, and continue to go deeper together into the Big Data culture as it is forming. As shared previously, a Big Data Champion, once established, could be a leader of this user group. In the interim, OmPoint can create and lead this group. For example, a weekly meeting to introduce a new Big Data technology, or use case (and how it was solved), would help educate and elucidate these issues.

5.9.2 Big Data Learning Lab: what to do.

In addition to a user group, an actual learning laboratory, with dedicated hardware and software would go a long way to allowing your employees to create a proof of concept (POC), and other experimental projects that encourage deeper learning in Big Data technologies.

5.10 Action Items for Section Five:

5.10.1 Review past decisions regarding on-premises vs. cloud hosting for continued viability of all the sub-categories mentioned in the road map from performance tuning to security.

5.10.2 Review past training effectiveness and decide upon 2018-2019 training goals.

5.10.3 Codify the Big Data business support structure.

5.10.4 Hold regular Use Case Discovery workshops internally, either lead by OmPoint or using internal resources.

5.10.5 Make final decisions on ingestion needs, such as PII and data masking, etc.

5.10.6 Finalize ingestion validation, testing plan, QA and UAT.

5.10.7 Embrace project management and communications processes that ensure more efficient work and clarity.

5.10.8 Implement the first Organization Big Data learning lab.

6 Security: don’t get hacked:

The purpose of this section is simply to remind all teams that security is an attitude and constant responsibility to maintain the safety of your organization and customer data. To this end, OmPoint will work with your IT team:

6.1 Make sure your organization has taken the following security related steps:

6.1.1 All traffic to and from Amazon S3 is encrypted.

This is a reference to "Amazon Simple Storage Service (Amazon S3)." Many people worldwide are using S3, which is an, "object storage with a simple web service interface to store and retrieve any amount of data from anywhere on the web."[15]

6.1.2 All traffic within the VPC uses a local S3 endpoint.

6.1.3 All buckets are restricted to access only from our account via the ACL.

6.1.4 All access lists are set to minimum necessary.

6.2 Action Items Section Six:

6.2.1 Immediately double check, with gap analysis to be performed by Laurent, that all the security layers are secure around ingested data. While we certainly trust the work of your organization's IT team, we don't want any unchecked risks to remain unresolved.

7 Disaster Recovery: It is bound to happen eventually.

Any Big Data road map will have considerations around disaster recovery. That said, each Big Data technology stack should be considered for the various features it has regarding failover, and related issues. In the case of Hadoop and Cassandra, and other technologies, we have what is known as a "Replication Factor (RF)." This means that all data written to the cluster is replicated for redundancy and the ability for the cluster to remain alive regardless of node failure. Furthermore, the replication can be made "rack aware" to further minimize the probability of failure affecting the cluster.[16] This all means that while Hadoop (and other tech) is resilient, if configured properly, we should have a policy and procedure around DR as it relates to Big Data.

7.1 Relevance to Big Data, policy and procedure:

We recommend that we create a meeting for OmPoint to understand your organization's policy and procedure around any IT DR issues, and based on that make a formal recommendation regarding DR as it relates to Big Data. There may be additional steps required, based on this formal review, or we may determine that, "all is well."

7.2 Backup:

From the point of view of simply backing up the organization's data which has already been ingested into your cluster, OmPoint will bring this issue up to your organization's IT dept.

Action Items for Section Seven:

7.2.1 Review current Organization DR policy and procedure and applicability to Big Data with to remain in compliance with agreed upon policy and procedure.

8 Governance[1]

Some of the issues around Data Governance are more obvious and others subtler. At your organization, while there is most likely already a notion regarding the need for proper Data Governance, the question remains, who is responsible for this, and what will be done?

8.1 Roles and Responsibilities

Two roles come up around Data Governance: Data Steward and a Report Steward. Initially these may be the same person, but the gist of it is that someone must do the work of Data Governance, and

that is the Data Steward. A Report Steward will make the effort, on behalf of the org, to create reports that tie all Big Data decision making back to real results (whether ROI, or any metrics which show the value added by Big Data). We will discuss the Data Steward role further below.

8.2 Auditing usage:

Any governance discussion will have a spotlight on auditing. Naturally, different tools have different audit capabilities, so this must be determined on a tool by tool basis.

8.2.1 For example, Datameer already has some features for this.

At the Datameer with Hadoop level there are two things that come to mind:

8.2.1.1 Enabling user click auditing in Datameer, whereby you can trace the clickstream of any user within the tool.[17].

8.2.1.2 Greater than this, and more fine grained is the EventBus Listener API, which allows for a more profound level of real time alerting and event management.[18]

8.3 Lineage: tracking data origins:

Being able to track the lineage of data from the source data set to the final analyzed data is valuable. As with auditing, each tool has its own capabilities in this regard.

8.3.1 Datameer has some features for this.

At the Datameer with Hadoop level we have the "Sheet Dependencies" which shows a GUI interactive diagram of how each Datameer worksheet depends upon a source. This is a click-through diagram which allows you to navigate the dependencies in real time.

8.4 Metadata management: data relevancy:

Managing the metadata (data which describes the data), is no small feat in Big Data. While OmPoint has started training about these subjects, there is a long way to go. One recommendation has been the creation of best practices regarding ingestion: column naming, column types, and whether to use "schema on read," or not. We are currently seeing vast data mapping and transformation rules being applied on a case by case basis. How the meta data being created is then stored long term has not yet been codified by many orgs. We recommend a meta data project be formed to tackle this issue in the medium and short term at your organization.

8.4.1 Needs a comprehensive metadata policy and procedure:

It would be wise to be consistent with the existing organization's metadata management policy and procedure, if there is one. If there is not, OmPoint can help shape that policy with you.

8.5 Data lifecycle management and policy enforcement:

All data is part of a life cycle, which can be understood as - ingestion to analysis, and then to archiving, or backup. To better understand this life cycle in a healthier way, we need to ensure proper handling at every step.

8.5.1 Ingestion:

As previously stated, we need to establish very clear ingestion guidelines, policy and procedure. For example, under what circumstances it will be permitted to use "Schema on Read" functionality, etc.

8.5.2 Retention

For example, in Datameer we have the ability to retain all the worksheets in a workbook, or just the final sheets which are most relevant to the final analysis. Doing this properly will affect the data footprint of the stored data in HDFS. While this comes up in Datameer training as a point, your

organization should publish clear guidelines about data retention in Big Data. This day to day retention of data should be carefully distinguished from archiving (backup) of data.

8.5.3 Archiving (Backup)

Data archiving or backup is a vital discussion which all Big Data organizations must resolve. There must be an existing policy and procedure about data archiving at your organization and this should be reviewed in light of Big Data. After this review, OmPoint can help with a recommendation.

8.5.4 Data Replication (Hadoop RF)

Data replication is already built into many of the Big Data systems using the replication factor in the administration configuration at the cluster level. For example, RF=3 means that there is one block which is the original data, and it is replicated two more times, so there are three blocks total. This is a standard setting, but it is configurable up and down.

8.5.5 Data Compression policy

OmPoint has been working with various compression strategies. The subject of how best to use data compression with Hadoop is a large one, and it is complex.[18] This should be dealt with as a separate project once the daily flow of work is stabilized. Ultimately, this falls into the category of Hadoop performance tuning. Other Big Data tools also benefit from data compression strategies.

8.5.6 Data stewardship and curation:

The Data Steward role requires a comprehensive understanding of your organization's data sources, ingestion and storage in Big Data tools, including but not limited to Hadoop. Some of the tasks and responsibilities are given below, but this is not meant to be an exhaustive list:

8.5.6.1 Data Catalogue and user access controls

As the Big Data investment continues to grow, your organization will need a "Data Catalogue," with strict user access controls regarding who is authorized to view, alter, delete and create new data. While this may be in place in some ad hoc way currently, a systemic policy and procedure needs to be in place for all Big Data usage.

8.5.6.2 How to manage Data Source Changes:

Recently the issue was raised regarding data source changes, and how that should be handled. What is needed is to codify what constitutes a "change," and how does that impact the work already done with that data source. In other words, if the data source changes, does that automatically mean the data has to be re-ingested into Hadoop, and etc. This is a deep and meaningful conversation.

8.5.6.2.1 What requires re-ingestion and re-testing?

Determining criteria of re-ingestion is one of the tasks of the Data Steward. Once re-ingested that data needs to be retested for validation and integrity.

8.5.6.2.2 What triggers a review based on a known Data Source change?

Some organizations have raised this issue in meetings, and there is an ad-hoc nature to how this is dealt with. What is needed is a clear policy regarding how data source changes trigger the actions of re-ingestion and testing. For example, we suggest the creation of clear rules such as: "Column name changes require re-ingestion of data." Or, "Data type changes require a re-ingestion of data," etc.

8.5.6.2.2.1 Table Level change, naming, adding or removing columns?

Further to this is the level of changes around adding, removing and changing entire tables. All in all, the subject of data source changes must be reviewed and a clear policy put into practice around this so that it never needs to be discussed in a meeting, but is always actionable.

8.6 Action Items for Section Eight:

8.6.1 Implement a comprehensive Big Data Governance program with the appropriate roles and responsibilities, and whenever possible engaging the vendor tools available. (For example, leverage the Datameer data governance features available in the most recent version of their product, etc.)

9 Repeatable process: fully independent

The final phase of any successful Big Data culture is to become fully independent of those who helped to put the tools and understanding in place. Like a child who matures into a young adult and differentiates from the parents, in the same way OmPoint aims to help your organization to mature to where little or no continued support is required.

9.1 Success means no need for OmPoint or vendor support:

While there will always be a need for technical support for Big Data tool bugs, and it is wise to have Big Data health checks with OmPoint, we envision a time when there is no need for Organization to have direct support from OmPoint. For this to take place and become a reality, the adoption of the points in this road map is a requirement, as is embracing a culture of change through continuous improvement.

9.2 What needs to be in place to support this for the long term:

To this end of long term success your organization must rely upon the Big Data key players we have previous mentioned, as well as real Change Agents in the company, who are fearless in their adoption of Big Data. The courage exemplified by these agents will help inspire and quell the fear in all the others who are in various stages of adoption. When this is supported by the executive staff, and the Big Data Champion, there will be success.

9.2.1 Change Agents apply here:

Any organization is susceptible to falling into unhealthy patterns of complacency, rigidity, stagnation, fear, and political self-sabotage. To help the Big Data culture at your organization thrive, Change Agents should be groomed and supported continuously. Big Data doesn't need more *yes-women* and *yes-men*, it requires creative problem solvers who find new ways of tackling Big Data problems with an array of tools.

9.2.2 Continuous improvement:

Living in a culture that supports honest feedback, healthy non-violent communication, as well as a deep desire for continuous improvement is vital to Big Data success. Gentle reminders about these issues will help foster such a culture, gradually, over the next few years.

9.2.3 Rinse and repeat: keep it clean.

In conclusion, there are many ways to adopt Big Data. We can see this from use cases and success stories around the world. Defense and other vertical sectors have been working with Hadoop and other Big Data tools for many years now. For every success story, there are many failed attempts which end in lessons learned. The goal of this Big Data Road Map is to apply the lessons learned from others, adopt the best practices already being lived successfully, to ensure that your organization maximizes the return on your investment in Big Data technologies. Let OmPoint help you achieve this level of success that we know is attainable.

9.3 Action Items for Section Nine:

9.3.1 Have a third party external assessment of how well your organization has adopted and lived up to this Big Data Road Map with OmPoint, and contemplate the report. Wherever needed, change to better live up to the stated goals of this program. Such an assessment should be conducted yearly, if not more regularly.

FOR MORE INFORMATION:

Laurent Weichberger
Big Data Bear, OmPoint Innovations
5116 Long Pointe Rd., Wilmington NC 28409
laurent@ompoint.com | ompoint@gmail.com
(928) 600-8898 mobile

End Notes:

1. See also: https://vision.cloudera.com/data-governance-in-hadoop-part-1/
2. From Jared Warren, Distinguished Engineer, Omnichannel Analytics and Insight at Organization (via email to Vikram and Santosh, and cc: Laurent on 11-16-2016).
3. Point and footnote removed. – LCW.
4. See: https://en.wikipedia.org/wiki/Column_family
5. See: https://neo4j.com
6. See: https://blog.cloudera.com/blog/2009/12/7-tips-for-improving-mapreduce-performance/ and http://blog.cloudera.com/blog/2009/03/configuration-parameters-what-can-you-just-ignore/
7. See: https://www.cloudera.com/documentation/enterprise/5-5-x/topics/cdh_ig_yarn_tuning1.html
8. See: yarn-tuning-guide.xls
9. https://documentation.datameer.com/documentation/display/DAS50/Tez+Tuning+and+Troubleshooting
10. See email from Tausif Shahid, to: Laurent, dated: 11/23/2016, subject: "Datameer REST API call to start workbook."
11. See emails from Gido Bauman, Datameer Tech Support team, November 2016.
12. See email from Adam Gugliciello, to Laurent: dated 11/17/2016, subject: ElasticSearch/Solr Integration. See also: https://www.elastic.co/products/hadoop
13. See: https://en.wikipedia.org/wiki/Marshall_Rosenberg
14. See: https://www.amazon.com/Anatomy-Peace-Resolving-Heart-Conflict/dp/1626564310 (2nd Edition)
15. See: https://aws.amazon.com/s3/
16. See: https://docs.hortonworks.com/HDPDocuments/Ambari-2.1.2.0/bk_Ambari_Users_Guide/content/ch03s11.html
17. See: https://documentation.datameer.com/documentation/display/DAS50/Logging+User+Clicks
18. See EventBus Listener API blog article by L. Weichberger: https://www.linkedin.com/pulse/datameer-eventbus-listener-laurent-weichberger
19. See: http://www.cloudera.com/documentation/enterprise/5-6-x/topics/admin_data_compression_performance.html

Appendix A: Organization Use Cases

In the following pages, I share some use cases which I have gathered while facilitating my Use Case Discovery Workshop around the world. We share them to show the usefulness of this exercise. Each team was asked to come up with a real-world organizational use case with the following criteria: Title, Priority (1=highest), Problem, Solution, Stakeholders, Reviewers, and Data sources. Enjoy.

Danny Loftis
AVP Data Integration at GM Financial
May 14, 2018, Danny was a client of Laurent's

Several members of my development team and I recently attended a training event to learn more about Spark and Scala that was led by Mr. Laurent Weichberger. This would be the 2nd "Big Data" week of learning for us to attend. The first week of training was held months earlier by a different provider. That training didn't stick very well, so we were understandably apprehensive. Our apprehension was immediately put to rest. Laurent has an obvious passion for the subject matter and a passion for sharing his knowledge with others. He was constantly engaging different team members to make sure everyone was getting the information. When he felt he wasn't getting through to someone, he took the extra time to make sure they got caught up before moving forward. His use of "real world" scenarios and actual professional experiences provided substance to the lessons, which gave us much more to grasp. At the end of the week, we didn't just feel like we learned something. We felt motivated to get started on the new material we just ingested. You cannot ask for more from a training experience. Sending several senior developers and managers to a week of training is a very expensive investment. Missing out on an opportunity to train with someone like Laurent is much more expensive.

Thank you, Laurent. We no longer look ahead at Big Data as a challenge, but as an opportunity. We have you to thank for that.

Anne Kwan
Software Engineer at Booz Allen Hamilton
October 27, 2017, Anne was a client of Laurent's

Laurent was the instructor for my Hadoop course, and it may be the best course I've ever taken! I really appreciated his candor, and he geared the course material towards our specific goals for the course. Additionally, Laurent kept the course interactive to help reinforce our learning and make sure we were engaged. The use case at the end was great to get our hands dirty.