Do Not consider building a Data Lake before reading this whitepaper
And if you have, consider adding one step to a better outcome

BigDataRevealed-VM allows us to rethink and re-architect Big Data Hadoop Projects/Implementations, by assuring lower risks of PII and Privacy Data exposure, removal of Anomaly Risks, while adding valued and needed metadata and cataloguing.

What is the current state of Big Data projects?

- Though the names of methodologies and vendors keep changing, and the number and sources of data have increased, the true core methodology, delivery and needs stay the same. The speed and value of delivery is key with cost a close second.
- We have legacy data sitting on Mainframes, RDBMS Servers with Oracle, DB2, Teradata, SQL Server, PostgreSQL, AS400, Word, Excel, PDF, XML and many more data types along with IOT (from websites, mobile devices, machinery, utilities, third parties and more).
- The goal today is to move as much data safely, cleanly and logically to new Big Data Environments to save money, hardware and Licensing fees. It is also assumed consolidating data will allow delivery of more meaningful and valued Business Intelligence, Predictive Analytics and Artificial Intelligence. In doing this we also need to absorb real-time streaming data and third party data to have a 360% view of our corporate, sales, customer sectors, accounting and forecasting.
- Just as we have done for the past 25 to 35 years, we still build staging areas, Operational Data Stores to prep the data, process ETL rules and validations, and acquire business rules to help understand this old, archaic data. This must be accomplished with the additional burden created by years of poor data practices, such as re-using database sections and columns, and allowing obviously erroneous entries to remain in the database. It's no wonder that Data Scientist/Analysts become confused, especially with most of the old data stewards, subject experts and documentation long gone.
- Unfortunately, today, many companies believe they can move all this disparate data into Hadoop or other Big Data NO SQL Databases and there will be tools available to solve their data problems, build metadata and catalogues while also identifying pesky Outlier/Anomalies and exposed PII/Privacy issues. Companies have discovered this is not the case and Data Scientists and Analysts ATTEMPT to eradicate and remediate this on the fly or build non-collaborative non-repeatable tools and processes themselves. This approach carries an extremely high risk of failure while consuming huge amounts of manpower and corporate dollars. Just how many successful Hadoop implementations are you aware of?

How did BigDataRevealed become a difference maker? And what does it do that other products and human intervention have been unable to do successfully and in a timely manner?

- We first began with an excellent Data Quality / Profiling / discovery and Metadata team of experts to provide knowledge and direction for a development effort that began from the bottom up. Every line of code was completed using only Hadoop ecosystem languages, Spark and MapReduce.
- We created Jar executables called by the BDR D3.js open graphical tools and a GUI Front-End using restful API's along with live streaming processes using standard HDFS, Hive, Hbase, Apache Drill and other Hadoop Eco-System and Framework Technologies giving our product the ability to live 100% within the Hadoop ecosystem and take advantage of its distributed processing / speed and data storage.
- Our Hadoop Framework API modules are callable from any third party or in-house process and can be seamlessly included as part of any ETL, BI, Predictive Analytics, AI or other process.
- We believe we are the only product that can make that claim. Using the Same Hadoop languages, we added Pattern Recognition modules that identify numerous PII and Privacy data formats, and then we had our developers build statistical modules capable of discovering Outlier/Anomaly data, all from within the Hadoop ecosystem using Pattern Detection, NLP, Data Mining, Deep Learning and more.

An almost unbelievable feature of BigDataRevealed is that a complete version of our product, BigDataRevealed-VM, comes preconfigured with a complete version of Apache Hadoop, and can easily be loaded onto a departmental PC or Laptop. Our code is so powerful and streamlined that it functions on a PC and still delivers all the metadata, cataloguing, discovery, pattern detection and outliers that are needed. You can pre-process departmental data before loading it into your production Hadoop environment, keeping your production Hadoop ecosystem pristine.

Hadoop, by its nature, strips off the catalogue and metadata values from incoming files, forcing Data Scientist to spend much of their time just trying to re-construct what had already been in place. BigDataRevealed has overcome that hurdle and easily accepts metadata and catalogue values when incorporating new files into the Hadoop environment. This capability extends not only to your production environment but also to your installs of BigDataRevealed-VM on Departmental PCs to eventually be used and ported into your company's primary Data Lake. Of course, BigDataRevealed contains many features to assist in further development of metadata and cataloguing information so that a truly rich description of your database will exist for use by Data Scientist, Analysts and others.

With NO cost for the first 30 days, you can use the BigDataRevealed-VM, fully configured out of the box with Apache™ Hadoop® delivering 100% of your needs to accomplish the following, again at NO software vendor costs.

1. Discover data patterns to determine the columnar metadata naming with a user participation Interface
2. Building a Cataloguing System, User and third party metadata
3. Identify Sensitive, Private Customer data allowing Isolation/Consolidation of these files
4. Discover Outlier/Anomalies that will skew and pollute the Data Lake and delivery results
5. Protect, notify you of what needs to be put in encrypted Zones, mask or eliminate this Privacy Data so to eliminate it accidently entering the Corporate Production Data Lake
6. Make available all the metadata for project collaboration across company groups, external consultancies, third party metadata tools to properly and consistently name file/columns to your company standards
7. Search for Banking Governance quicker within smaller departmental or group staging areas for violations that can be more quickly identified and remediated prior to polluting and getting lost in the Corporate Data Lake costing risks, fines and the nightmares of this data being hacked
8. Analyze and properly Name Folder and sub folder semi-structured and unstructured data such as word, excel, xml, pdf, email, resumes, rtf and many other file formats with BigDataRevealed's discovery of these data types and suggested Columnar Naming.
9. Process live streaming feeds from utilities, banks, manufacturing, retail transactions and more.
10. Set-up the proper data stewards and subject matter experts to be warned and notified of anomalies and violations needed to be acted on immediately to limit the time of these exposures from hackers and auditors.

Once you have this model and methodology in place and monitor this process over time, then just add this Node to your Corporate Data Lake or port the Data into your Corporate Data Lake along with the associated Metadata and Cataloguing information, assuring a much cleaner transition of your departmental or groups information into the Corporate Data lake.

BigDataRevealed can, within minutes, be installed in your Corporate Data Lake Hadoop environment to continue protecting your Data Lakes integrity and reliability and reduce risk of exposures if or when hacked or audited.

WE do offer training, onsite and offsite services to expedite your efforts, training of your trainers, training of your third-party consulting companies of your choice and will even fix bid detailed defined projects.

After download, use our wizard to install our VM product on your department PC or laptop. http://bdrvmware.bigdatarevealed.net/bdrvm/BigDataRevealedVirtualMachine-Quickstart-v1.1.ova

Steven Meister – 847-791-7838 steven.meister@bigdatarevealed.com –
http://bigdatarevealed.com/video-links-vm-download