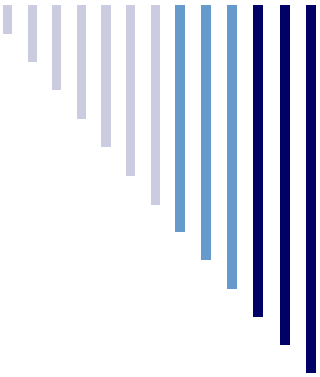

A QuantUniversity Whitepaper

Big Data Readiness

5 things to know before embarking on your first Big Data project

By,
Sri Krishnamurthy, CFA, CAP
Founder
www.quantuniversity.com

Summary:

Interest in Big data technologies has significantly grown in the last few years. Data growth has exploded in the last decade. Information systems are generating huge amounts of data, trading data is becoming much more granular and at higher frequency, unstructured data from Twitter feeds, blogs etc is becoming mainstream and technologies to produce, transmit, store and process these massive datasets is collectively enabling the Big data revolution. Data science is the one of the hottest areas for internet startups. Cloud computing is making access to unlimited hardware at your fingertips. Significant venture capital money is flowing to support newer and innovative technologies that leverage vast amounts of data to create platforms and applications that could enable decision making close to the proverbial "speed of light". Technology industry stalwarts like Google, Microsoft, Amazon and Oracle are making significant investments in Big data technologies. The applications and opportunities seem endless and the promise of revolutionary applications in healthcare, finance, space research and pure science has led various industry and academic organizations to significantly ramp up efforts to develop technologies to help analyze massive data sets. Quants have been in the forefront in the financial industry in adopting revolutionary technologies and the Big-Data phenomenon has undoubtedly caught interest in the quant community. In every quant gathering, there is at least one reference to Big data and discussions on the potential and possibilities. Quants have started to frequent Big data conferences and events to know more about technologies and opportunities in this space. I have had multiple discussions with clients who want to start Big data projects but many still struggle to understand what it is about and how they can leverage these technologies to further their quantitative research ideas. The information overload on Big data and the umpteen numbers of sources of Big data, primarily vendor and media driven, is adding to the confusion leaving quants to ponder on whether Big data is another fad or is there a real opportunity they should try out to gain an edge. I started out to write an article to introduce Big data but rather than just rehash information that is widely available, I thought it may be useful to share the my perspective as a quant practitioner with experiences from the analytics and the quant worlds on pointers to help quants understand the realities before embarking on a Big data project. In this article, I will start out with a brief introduction to Big data and point you to sources where you can learn much more about Big data. I then discuss five key things quants should consider before embarking on your first Big data project. I will then conclude with pointers on how quants can keep themselves in tune with the rapid innovations happening in the Big data world.

Big data: What is it about?

One of the primary questions that need to be asked is what is Big Data? Recently I had a conversation with a quant and his IT manager from a large financial institution. The conversation went something like this.

Manager at a large financial institution: We are looking to leverage Hadoop for our datasets. Can you propose a Big data solution?

Sri: What is "your" definition of Big data?

Manager at a large financial institution: We have tons and tons of data in our databases. We need to analyze this to see if we can make sense of it.

Sri: Why do you think you need a Big data solution for your use case?

Manager at a large financial institution: Well, our current systems can't scale to process all this data. I suppose a "Big" data solution could help..

I am sure you may have heard of similar conversations in other contexts. My point is there isn't a clear definition of what Big data is. One of the most comprehensive reports I have seen was published by McKinsey [1]. In this report, they refer to Big data as datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. Though the definition is generic, it reflects the current thinking in the expert community that the Big data field is still nascent but rapidly evolving and influences from various communities including statistics, data management, economics, machine learning etc. is shaping how Big data solutions would manifest in the future. In addition, Big data is characterized by four tenets, Volume, Velocity, Variety and Veracity generally referred to as the four V's of Big data [2] and summarized in Fig 1.

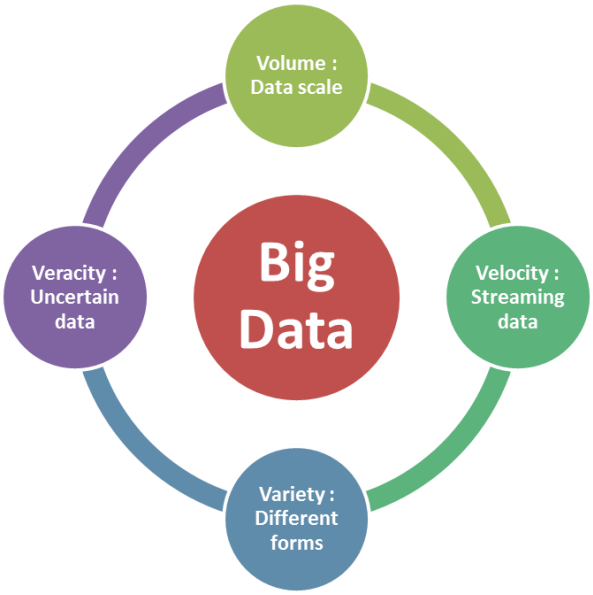

Fig 1: The Four V's of Big data

So if traditional data management software cannot deal with Big data problems, what other solutions are available? In the last few years, there has been significant progress in Big data technologies notably in the open source realm. The Apache foundation's Hadoop project [3] enables distributed computing across large number of clusters. In addition, Hadoop related projects like Cassandra, HBase, Hive, Mahout, Pig etc. has provided an ecosystem of technologies that enable scaling, storing, summarizing, data mining and parallelizing for Big data problems. In addition, many vendors such as IBM, Microsoft, Amazon, Google, Oracle etc. are either supporting the Apache stack or providing their own alternatives for Big data problems. Rather than summarizing the entire literature on Big data, here are a few books and links that would help enhance your understanding on Big data. For an introductory reading, Mckinsey's report[1] is a great place to start. Mayer-Schönberger and Cukier's book[4] provides a comprehensive introduction to the current state of Big data and the possibilities it promises. Rajaraman and Ullman's Mining of Massive Datasets book [5] provide a deep-dive on some of the typical problems related to Big data and the state-of-the-art algorithms that are out there to solve these problems. Tom White's Hadoop Book [6] is a great place to start when experimenting with your first Big data project. In addition, there are various courses available on Coursera and Udacity to help enhance your understanding on Big data. BigdataUniversity[7] also provides self-paced courses to on various big data technologies. Even though it may not be feasible to be an expert on all these technologies, having a general awareness and knowing whom to contact for expertise would help plan a successful Big data project.

Things to know before embarking on your first Big data project:

At QuantUniversity, we have consulted for various large financial institutions exploring Big data technologies. The use cases range from preliminary exploration to full-fledged deployment of Big data technologies to achieve specific use cases. Working through projects, we have seen our clients repeatedly face challenges when working on their first Big data project. Typically, a significant education component is involved to help our clients understand their choices and the pros and cons of specific approaches. Sometimes, the expectations for the project may not be realized if there are gaps in understanding on what Big data technologies can do. In this section, we highlight five things to know before starting your first Big data project.

1. Do you really have a Big data problem?

Marketing by various vendors and the media has turned Big data as a panacea for most problems. In every conference I have been to, vendors talk about possibilities in generic terms but I have heard very few financial institutions discuss specific problems and how Big data solutions helped solve these problems. The unfortunate reality is that Big data technologies work best for only a certain classes of problems. The primary question you need to ask is whether you really have a Big data problem. Financial institutions typically work with large amounts of very structured data. It is enticing to embark on a Big data project to "explore possibilities". However, if the problems you currently face in analyzing your datasets isn't clearly identified, it may be a less rewarding and sometimes futile to deploy Big data technologies to such problems. We advocate defining quantifiable metrics for your problems and then seek solutions that would help attain these metrics. The solutions may be in the traditional technology realm or in the Big data space. But having a clear notion of what the problem is and the ideal solution would help avoid wasted efforts in trying out solutions that may not be optimal for your problems. This is not to discourage you. At times, we have seen organizations gain significant advantages by using Big data technologies. However, for most cases, we advocate running your first experiments in a prototype mode till you are absolutely sure that the solution meets your criteria before embarking on a large scale Big data project.

2. You have a Big data problem. What dimension of the 4V's should you focus on?

Typically, when you hear about Big data problems, you hear about the terabytes and zettabytes of information that is produced and the challenges in sourcing, storing and analyzing these huge volumes. However, Big data problems are also characterized by the velocity, variety and veracity dimensions. For example, to analyze vast streams of tick data in a trading house, the velocity is more prominent compared to other dimensions. Handling text data in combination with structured numeric data requires different data designs. Handling massive streaming data requires different architectures when compared to handling massive volumes of data. Also, open source is a huge driver for innovation in Big data and technologies are in different stages of maturity. For example, Hadoop has been in development for more than five years, whereas frameworks like Storm [10] for streaming data was released initially in 2011. We advocate clearly setting goals for your problem you are trying to address and explore technologies that could help with your problems. Once the relevant technologies are identified, particular attention should be paid to analyze the maturity, community support and backing from other players in the ecosystem to ensure that a solution developed relying on a chosen Big data technology can be maintained and supported in the long run. Organizations must also introspect to see if they have an appetite for open source technologies and if they are ready to use these technologies as early adopters.

3. *Experiment to get a comprehensive picture on what it takes to implement a Big data solution*

Big data technologies are unique in that there are significant cross-disciplinary aspects involved in the design of a solution. Moving away from traditional databases involves new ways of representing the data. A Big data system is typically deployed on multiple clusters and or the cloud that would require significant knowledge in setting up the system infrastructure. In addition, aspects that are typically ignored in locally deployed solutions like security, network bandwidth and latency are to be considered when designing a Big data solution. Also, cloud vendors charge based on usage and therefore there is an additional consideration on how much compute capacity needs to be sourced for these problems. For organizations trying out Big data solutions for the first time, we typically advocate to experiment with a small problem to get a comprehensive picture on what it takes to implement a Big data solution. This would help engage all the stake holders to ensure that there is a collective decision in case a larger scale project is to be worked on.

4. *A Big data demo or a prototype is not "the solution"*

Just because a demo or a prototype works, it doesn't mean it is a solution to your problem. Like any other data driven solution, exhaustive testing and validation must be done to ensure that the solution is stress tested enough to be deployed into production. We have seen cases where during the testing and prototyping phase, only a portion of the data was used but the production data was significantly different than the test data. The outputs must be scrutinized and monitored to ensure that the system is performing as per the metrics defined. In addition, if open-source technologies are used, support and contingency plans must be developed if the solution is going to be a critical piece in the decision making process. In addition, if any of the dimensions of the 4 V's change, the system must be tested to ensure that the system still performs as desired. Deploying a Big data solution is not a one-time operation. Organizations must plan to be actively engaged in ensuring that the system performs as per design on an ongoing basis and structures need to be put in place to support such a dynamic system.

5. *Seek expert guidance*

Embarking on a Big data project requires significant co-ordination between various stake holders in the organization. IT departments have the most know-how to assist in the infrastructure setup for such projects. If a cloud environment is considered, security and network issues need be addressed and IT departments should be engaged early on to ensure that all stakeholders are in agreement. In addition, risk departments should also be involved to ensure that any data that may be sourced from outside or sent outside for processing don't violate any of the risk policies of the company. Large organizations with sophisticated IT and software development groups typically have in-house expertise to assist in taking on or setting up Big data projects. However, if your organization doesn't have the in-house expertise, we advocate seeing expertise from outside either through vendors, consultants or by hiring Big data specialists. In many of the projects we have done for clients, education has been a significant piece of the advisory. Seeking expert guidance would help organizations plan judiciously and maximize the outcomes of investments when embarking on large scale Big data projects.

Conclusion:

We understand that planning your first Big data project is a challenge. The Big data references we provided earlier will help you get started. Considering the field is evolving rapidly, there are many new books and articles that being published regularly. The only way to keep up with the innovations in this space is through active engagement and participation. We encourage you to explore Meetups and other Big data events in your communities. There are also many online webinars and conferences hosted in a virtual environment. Videolectures[8] is one such forum where high quality presentations from top researchers in the world are made available for free. Additional presentations on the topic and pointers to enhance your learning are available on QuantUniversity's website [11]. There are various linkedin groups dedicated to the topic and blogs such as the Smarter Computing Blog [9] provide updates on the latest innovations in the space. Embarking on your first Big data project can be a challenge and getting all aspects to successfully launch your first project could be daunting. With active engagement, education and careful planning, companies can tackle this challenge and gain the technology edge that was never possible just a few years ago. We are seeing the beginnings of a huge revolution and we are bound to see many innovative applications enabling new products and services in financial services organizations in the near future. Stay tuned for the upcoming Big data revolution!

References:

1. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
2. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg
3. http://hadoop.apache.org/
4. http://big-data-book.com/
5. http://infolab.stanford.edu/~ullman/mmds.html
6. http://hadoopbook.com/
7. www.bigdatauniversity.com
8. http://videolectures.net
9. http://www.smartercomputingblog.com/category/big-data/
10. http://storm-project.net/
11. http://www.quantuniversity.com/

QuantUniversity offers quantitative modeling and consulting services to financial institutions and specializes in analytics, optimization and big data solutions. Sri Krishnamurthy, CFA, CAP is the founder of www.QuantUniversity.com , a data and quantitative analysis company. Sri has significant experience in designing quantitative finance applications for some of the world's largest asset management and financial companies. He teaches quantitative methods and analytics for MBA students at Babson College and is the author of the forthcoming book published by Wiley titled "Financial Application Development: A case study approach". Sri can be reached at sri@quantuniversity.com

This article will be published as a column in the March 2014 edition of the Wilmott magazine

QuantUniversity, LLC

A data and quantitative analysis company

51 Pearl St, Unit 206 Malden MA-02148 | **Ph** 617-283-7904

www.quantuniversity.com | Sri@quantuniversity.com