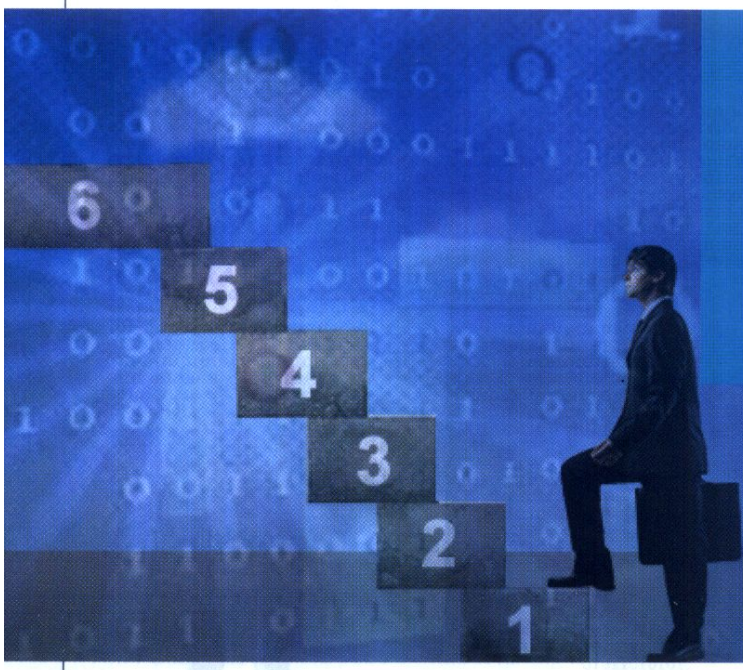

Six Key Steps in Framing the Right Big Data Strategy

As data volume increases, the ability to collect and present it in a way which the business can understand to make decisions faster than the competition will be the key to keeping business competitive. Here we list six key steps in framing the right strategy

- Yogesh Dandawate, Senior Architect, Analytics and Allied solutions, CloudMoyo

Enterprises today are being hit hard by the data tsunami and face an uphill task to ride this enormous data wave and turn it to their advantage. The market expects enterprises to be ready with smart strategies to deal with the information overload and consequently organizations now have to invest in time and resources to devise an effective big data strategy to harvest the large amounts of available data and unlock the hidden value buried in that data.

Key Considerations for a Big Data Strategy

As data volume increases, the ability to collect and present data in a way which the business can understand to make decisions faster than the competition will be the key to keeping business competitive. However there are several challenges to be dealt with before such data driven decision system is made available to the business. Some key considerations for an effective big data analytics strategy are:

1. Investing in the right talent

According to analyst firm McKinsey &Company, by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions. Similarly in 2012, Gartner analysts predicted that by 2015, 4.4 million IT jobs globally will be created to support big data with 1.9 million of those jobs in the United States.

There is a growing community of developers who are upgrading their knowledge of tools comprising the Hadoop ecosystem. Despite the number talk, the reality is that there is a dearth of skills in the market. It is essential to identify right candidates and train them to gear up to data challenges. People investment is key for analytics strategy.

2. Selecting the right big data technology platform

There exist several big data platforms, selecting the right one to suit ones organizational needs remains a challenge. In our experience, we found that most organizations start their discussion about big data with Hadoop, which although is a de-Facto standard for processing large volumes of heterogeneous data, can be an overkill for certain needs. In some cases a simple data warehouse can solve the problems effectively. Organizations where data is manageable, structured and static our recommendation has been traditional BI systems to solve the operational, performance and forecasting needs of the organization. But if the long term goal is to deal with unstructured and social data, organizations should plan Hadoop based systems. There are several technology options that are available today- the decision of selection is driven by:

- The future plans of the organization and current needs
- The technology that the organization is currently using
- The scale of data and kind of storage need, whether information will be stored on premise or over the cloud
- The type of data that needs to be analysed - static or real time, structured or unstructured

Yogesh Dandawate
Senior Architect, Analytics and Allied solutions, CloudMoyo

"Big Data is of limited value if users cannot understand the analysis. Systems with rich variety of visualizations become important in conveying to the users the results of the queries in a way that is best understood in a particular domain."

- Type of visualizations needed and compatibility of visualizations with the platform
- Platform should be able to manage Data Integrity, Data Governance and Data Privacy challenges

3. Synchronized, scalable data discovery techniques for handling heterogeneous sources

When humans consume information, a great deal of heterogeneity is comfortably dealt with. However, for machine analysis the data needs to be homogeneous and uniform. The data needs to be carefully structured or a transformation needs to be applied over the data to make it uniform. Although big data platforms promise to manage large scales of data, the initial heavy lifting has to be done by organizations- they have to pull data into the platform from various sources. Data from heterogeneous systems needs to be handled efficiently and has to be churned and normalized before it can be used for further analysis. This requires building interfaces to fetch the data into the big data platform; most of these platforms are equipped with adapters that help you connect with your existing data stores. The key decisions that need to be made are – which data needs to be extracted, how much data should be fetched, what is the frequency of data refresh. For accurate analysis the data should always be in sync with the data sources and should have the capability of scheduled refresh.

4. Selection of right visualization tools

Big Data is of limited value if users cannot understand the analysis. Systems with rich variety of visualizations become important in conveying to the users the results of the queries in a way that is best understood in a particular domain. Right visualization of data helps in quick understanding of data and fast decision making. A plethora of visualization tools are available in the market for this purpose. Selecting the tool best suited for your kind of data is key to utilizing the knowledge derived from analysis efforts.

5. Data identity and security management

As big data becomes more user-friendly, concerns around secured access to sensitive data are inevitable. Data security does not have to be compromised when an enterprise undertakes big data initiatives. To run big data analytics, large data sets are split up into more manageable portions and are then processed separately across different Hadoop clusters. They are then recombined for desired analytics. This process is highly automated and involves great deal of machine-to-machine (M2M) interaction between the clusters. Hadoop infrastructure has several levels of authorization for access to the Hadoop cluster, inter-cluster communications and cluster access to the data sources. Because big data means increased access to sensitive information, organizations must take proactive measures to roll out a comprehensive and consistent identity and security management strategy.

6. Big Data governance

Governed data is reliable, secure and ready to use, while data from an ungoverned terrain has little value for big data analytics. A well-thought out data governance policy can enable organizations to increase the value of their information. It can help secure company's legacy knowledge and intellectual property. A data governance policy can help comply with regulations, e.g. Regulations such as the Federal Rules of Civil Procedure (FRCP), the Federal Rules of Evidence (FRE), the Health Insurance Portability and Accountability Act (HIPAA), Sarbanes-Oxley (SOX), and others.

Today it is critical to have a data governance policy to outline who has access to what kind of data, who owns which components of data, what is the data retention policy, what should be the disaster recovery policy, etc. The plan should also have the statistics on rate at which data is growing and cost of maintaining such large volumes of data. This is a critical consideration when data storage is cloud based.

View from the trenches

Having a comprehensive organization big data strategy is key to valuable outcome - strategy should encompass not only data discovery, storage, analytics and visualization aspects, it also should have a detailed plan on facets of data governance, identity and security management. Having all these critical ingredients in place will guarantee smooth sailing even in rough data seas. □