**Big Data 2.016 - Practical implementation challenges**



We have been witnessing the journey of “Big data” from inception until 2015. Big data got a fair share of attention and has passed through the hype cycle. In 2016, there is a different set of challenges that organizations need to overcome for realizing the true potential of Big data through analytics. These challenges are primarily related to implementation of Big Data projects.

**Until 2015**

By the end of 2015, most organizations would have already invested in some pilot Big data infrastructure (appliance, datacenter, hadoop, NoSQL store) and might have executed a few proof-of-concept projects. These projects would have helped provide an idea about the complexity and investment required for implementing Big data projects.

**2016 & beyond**

2016 marks the phase of implementing advanced analytical solutions in the Big data infrastructure. This is the time when the business value from Big data is realized. Below are few of the challenges that would be part of (at least) the 2016 journey as the implementations become real.

* **Infrastructure**

As implementations are productionized, there arises a need for separate environments for development, staging / testing and production within enterprise systems. The challenge would be to acquire, plug and commission the right infrastructure to roll out within the organization. There is a fair amount of estimation required to determine the hardware (nodes / racks) based on the workload to be supported in the longer term. Along with the hardware, determining the appropriate versions of operating systems, VMs' and tools might take time and investment. Once the infrastructure and tools are determined, it needs to be integrated within the data centers and other existing enterprise systems.

* **Security & Governance**

There are mechanisms available for securing the Big data environment at various levels - network, storage encryption and access. Integrating the security tools with the enterprise user directory and authentication protocols is challenging due to the maturity of the said security tools. Also, before we begin the ingestion of data in the environment, it is important to define the governed areas for hosting data per the type - raw, mapped, processed, production runs and so on. The defined zones need to be created with appropriate policies based on user / team and application access.

* **Data availability**

Making data available in the data lake for analysis has been one of the biggest challenges. Almost all analytic use cases use enterprise data. With the polygot persistence pattern, pulling data from system of record, data warehouse, raw files and MPP systems is a daunting task. Ingesting data from multiple sources / teams requires a lot of coordination and effort for project management to ensure smooth execution. The effort might require creation of external views in existing systems to enable data pull in the cluster using ingestion tools. Also, if the analysis require maintaining a bi-temporal view of the data (snapshots), it becomes much more difficult if the source systems do not have history of change or audits to maintain changes.

* **Processes**

Execution of Big data projects involves multiple teams (IT, data engineers, data analysts, data scientists, business owners) . Processes and handshakes are required to streamline project execution and avoid project delays. Execution workflows need to be defined for project initiation, data source pull, model development standards, data sharing and operationalization. The organization structure within an enterprise drives the processes being created and owners of those processes.

* **Standardization**

Data being available from disparate sources calls for standardization across data sharing, date formats, contact number formats, mapping logic, etc. There should be standard ways to for identifier sanity checks, data quality checks, model construct development, scheduling and operationalization.

* **Consumption**

Big data analytics motivates near real-time consumption. The traditional file sharing of scores is no longer preferred. Applications require access to analytical outputs as they happen and also on a request basis. Based on the needs, new consumption patterns are defining different mechanisms - FTP, RDBMS store, exposed ReST services (pull) or enterprise service messaging systems (push). These consumption patterns bring along new integration challenges for Big data analytical outputs.

* **Skill sets**

Big data tools are driven by open source development. This makes the tools and technologies very dynamic, with new versions and features launched frequently. Organizations need to train internal teams to adopt to newer paradigms of analytics development like distributed computing and near real time computation.

Big data analytics is becoming an integral part of decision making across various enterprises. Despite the challenges, implementation of Big data analytics is believed to be the key success criteria for organizations. It has always been about the most important V - ‘V’ALUE.