Seven Questions to Ask Before Investing in New or Additional Big Data Infrastructure for Hadoop

Enterprises have several infrastructure options to choose from when it comes to running Hadoop jobs. Before making a Big Data infrastructure investment for an upcoming project (or one that’s causing you angst), ask yourself the following questions to hone in on what’s best for your business.

1. **How many distinct Hadoop clusters do I need to run?** In addition to needing separate clusters for development and production purposes, will you need to provide diverse stakeholders in your organization the opportunity to run data on their own clusters? If so, look for infrastructure solutions that give you the ability to share a single, unified infrastructure to run multiple clusters. This will enable agility and flexibility for your users while ensuring higher utilization from your hardware investment.

2. **How often will I run Hadoop jobs: 24x7, weekly, monthly, or quarterly?** When jobs are not running your expensive physical systems will be idle, so building an infrastructure that permits you to instantly scale up or down, as needed, will help control costs without sacrificing service.

3. **What is the process for running jobs on data or files on my existing storage systems like NFS?** Most infrastructures require you to copy data into a HDFS file system before your Big Data applications can access it. This time-consuming, expensive process results in the creation of multiple copies of data. Choose infrastructure that gives you the option to access and run Big Data jobs directly from your enterprise storage.

4. **How will I implement security and multi-tenancy above and beyond kerberos?** If a Big Data job requires the use of sensitive data, be aware that some infrastructure options require you to copy and move that data from your existing enterprise-class storage systems into HDFS before it’s accessible to Hadoop jobs. This results in multiple copies of data to secure and manage. Ensure your infrastructure does not require you to copy and move sensitive data off your organization’s secure servers.

5. **What speed do I require for my Big Data jobs?** The speed at which a Hadoop job runs depends on its ability to be broken into smaller data sets for simultaneous processing. The number of nodes in a cluster limits this “parallelism.” For a highly parallelizable job, you want as many nodes as possible to quickly complete the job. However, when running a less parallelizable job, you don’t want to pay for unused nodes. Choose an infrastructure that allows you to easily adjust job compute resources for maximum cost efficiency.

6. **Do I have in-house Hadoop infrastructure expertise?** To achieve maximum performance, Hadoop clusters require careful, precise tuning tailored for specific applications. Keeping Hadoop software current with patch sets and functional improvements also requires specialized expertise. Opt for an infrastructure that simplifies the complexity of Big Data management by providing end users with self-service provisioning and auto-tuning of parameters based on cluster size and resources.

7. **Is my Hadoop infrastructure future proofed?** Applications written for Hadoop constantly change, and chances are the ones you run today will not be the ones you’ll use a year from now. Ensure your infrastructure does not lock you into a specific vendor’s Hadoop distribution or a limited set of currently available applications.

---

**Find the Hadoop Infrastructure that’s right for you.**

Hadoop jobs can run on several types of infrastructure – from physical in-house cluster(s) to the public cloud, to virtual clusters in an in-house private cloud. BlueData, a pioneer in Big Data private cloud, streamlines and simplifies Big Data infrastructure.

**Test drive EPIC, our Big Data infrastructure solution today.** Learn how »

---

**About BlueData**

BlueData is the pioneer in Big Data private cloud. The company is democratizing Big Data by streamlining and simplifying Big Data infrastructure and eliminating complexity as a barrier to adoption. With its EPIC software platform, enterprises can now build agile, secure and cost-effective Big Data deployments that deliver value in days instead of months and at a cost savings of 50%-75% compared to traditional approaches. BlueData, enterprises of all sizes can create a public cloud-like experience from their on-premise environments and get the same value out of their Big Data as companies like Google, Facebook and Yahoo at a fraction of the cost and with far fewer resources. Based in Mountain View, CA, BlueData is founded by a highly experienced team from VMware, Akamai, Intel and SGI and backed by industry luminaries from Silicon Valley. For more information, visit bluedata.com or contact info@bluedata.com.