Hadoop-only deployments using the open-source version from

Apache. The open-source version of Hadoop is being used for an

amazing array of applications

* Put the Right Infrastructure in PlaceThe Hadoop framework works on the principle of moving computing closer to where the data resides, and the framework typically runs on large server clusters built using standard hardware. This is where the data is stored and processed. The combination of Hadoop infrastructure with standard server platforms provides the foundation for a cost-efficient and high-performance analyticsplatform for parallel applications.The Cost of Big Data Analytic.
* hardware and software choices made at design time can have significant impact on performance and total cost of ownership.

**Setting up hadoop System Architecture**

Each cluster has one “master node” with multiple slave nodes.

The master node runs NameNode and JobTracker functions and

coordinates with the slave nodes to get the job done. The slaves

run TaskTracker, HDFS to store data, and map and reduce functions

for data computation. The basic stack includes Pig and JAVA for

language and compilers.

Cluster hardware setup:

-Number of nodes in cluster. Need One master node and number of slave nodes based on project .

* Each server capacity
* Each node disk requirements
* Each node processor requirement
* Each node memory requirement
* OS requirements : Linux or Windows
* Hadoop Frame work v0.20.1

Configuration requirement :

* Java versions prior to 1.6 will not support all of the language features that Hadoop Core requires. In addition, Hadoop Core appears to run most stably with the Sun Java Development Kits (JDKs); there are periodic requests for help from users of other vendors' JDKs. The examples in later chapters of this book are based on Hadoop 0.19.0, which requires JDK.

**New requirements for this project:**

. Will only one or a few source systems load Hadoop, so that you can just use the HDFS API, or

will there be a significant number of source systems?

* Network architecture
* Operating System
* Hardware requirements
* Hadoop software installation/setup

Dedicate TOR switches to hadoop

Use dedicated core switching blades or switches

Ensure application servers are “close” to hadoop

Ethernet bounding for increasing capacity

Basic hardware recommendation:

Namenode/ JobTracker (2\*1GB/s Ethernet, 16 GB of ram, 4xCPU, 100 GB disk)

Datanode ( 2x1Gb/s Ethenet, \* GB of ram, $xCPU , Multiple diswks with total amount of 500+GB).

Do not mix up hardware when building up Hadoop cluster.

so to avoid this, somebody needs to know where Data Nodes are located in the network topology= give it a name.

Understanding the scope

* Enterprise data integration . Perform processing of routine data integration tasks and manage metadata with an integrated development environment (IDE) for standardization, reuse, and productivity improvement.
* To unleash the power of Hadoop, you need to assess where and how to perform data integration tasks involving Hadoop. You can ask the following questions to select your approach:
* Will only one or a few source systems load Hadoop, so that you can just use the HDFS API, or will there be a significant number of source systems?
* Hadoop integrated with traditional databases. Organizations with traditional data warehousing and analytics in place have the option of extending their existing platform to include an integrated Hadoop implementation. Connecting existing data management resources to Hadoop provides an opportunity to tap both structured and unstructured data for insights. F

Informatica [PowerCenter version 9.1](http://www.informatica.com/products_services/powercenter/Pages/index.aspx) includes connectivity for HDFS, to load data into Hadoop or extract data from Hadoop.