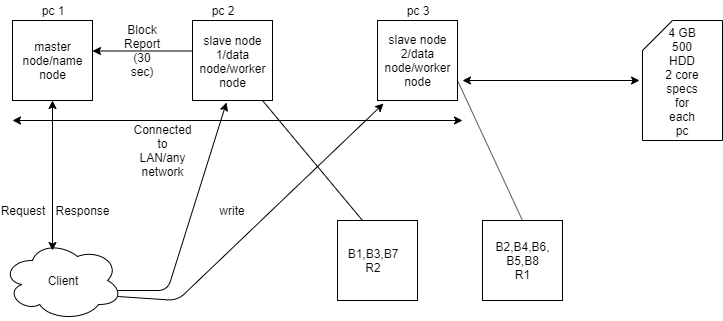
* Overview: (must know)
  + Linux
  + Sql
  + Python
  + Cloud computing
* Distributed Computing:

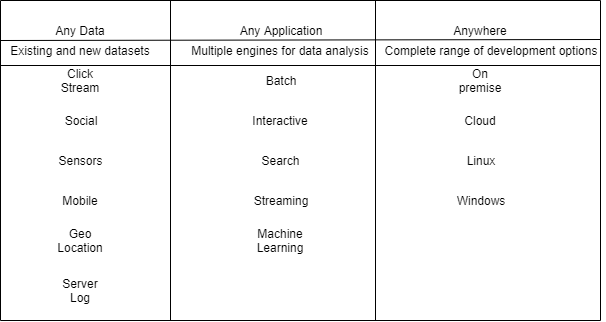


* + Specs: 16GB RAM , 1 TB storage , 4 core
  + When connected together and feels like each computer has power of all computers available.
* HDFS (Hadoop Distributed File System) :
  + It uses horizontal scaling.
  + Advantage : 0 downtime.
  + Can scale up or scale down according to need or demand.
* BIG DATA : (where did the phrase big data come from and what does it mean?)
  + The term big data came from computational science.Specially it is used to describe a scenario where the volume and the variety of the data overwhelm the existing tools to store and process it.
  + In 2001 the industry analyst DOUGH LANEY described big data using the three V’s. (Volume , Velocity , Variety)
  + Volume :
    - It refers to the magnitude of the data that is being generated and collected.It is increasing at a faster rate from TB to PB.
    - Challenges:
    - Storage cost
    - Filtering and finding relevant and valuable information in large quantity of data that often contains much information that is not valuable.
    - The ability to analyse data quickly enough in order to maximize business value today and next just next quarter or next year.
* Variety :
  + It refers to different types of data that are being generated and captured.They extend beyond structured data and fall under the category of semi-structured and unstructured data.
  + Structured data : The data that can be organized using a predefined data model are known as structured data. (eg : RDBMS)
  + Unstructured data : The data cannot be organized using pre defined data models. (eg : Video,Text,Audio)
  + Semi-structured data : The data that fall between the category of structured and unstructured data. (Partial data model) (eg : XML,Web Pages)
  + Challenges :
    - How to gather,link,match,clean and transform data across systems.
    - How to connect and co relate data relationship and hierarchy in order to extract business value from data.
* Velocity :
  + It refers to the rate of generation of data. (eg : Data generated through mobile apps)
  + Challenges :
    - Not reacting quickly enough to benefit from the data(eg : Data could be used to create a dashboard that could warn of a security breach or machine failure).

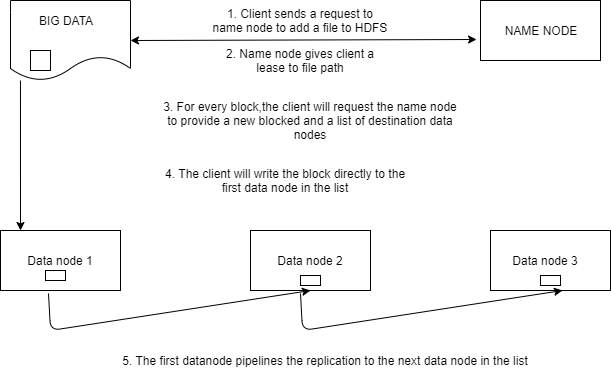


HDFS ARCHITECTURE

* Details about diagram :
  + B=block and R=replication
  + Secondary node >> master node (metadata)
  + Block report also known as heart beat.
  + Replication factor (default value : 3) (configurable)
  + File size(convert in MB) / blocksize = value
    - Where blocksize(default : 128) (configurable)
    - Value obtained is known as chunking of data.
* Apache Hadoop : (what is apache hadoop?)
  + It is a scalable,tolerant,open source framework for the distributed storing and processing of large sets of data on commodity hardware(daily used laptops or desktops).
    - Scalable :
      * Hadoop cluster can range from as few as one machine to 1000s of machine(using horizontal scaling).
    - Tolerant :
      * Hadoop services become fault tolerant through redundancy (eg : the hdfs file system automatically replicates data blocks to 3 separate machines assuming that node cluster has at least 3 machines in it ).
    - Open source :
      * Hadoop development is a community resource governed under the licensing of the apache software foundation.Anyone can help to improve hadoop by adding features,fixing software bugs etc.
  + Benefits :



* HDFS :
  + It is a java based file system that provides scalable and reliable data storage.
  + Key concepts of HDFS includes :
  + Write once and read many times (worm).
  + Divides file into blocks and distributed across the cluster.
  + Store multiple replicas for each block for reliability.
* HDFS Components :
  + The namenode is the master service of HDFS.It determines and maintains how the chunks of data are distributed across the data nodes.A namenode represents a single name space.
  + Data never resides on a name node.
  + The name node performs tasks by maintaining two files.
  + fsimage\_N : contains entire file system,(name space including the mapping of blocks to files and file system property).
  + edits\_N : the transaction block that returns every change that occurs to file system metadata.
  + Data nodes store chunks of data and are responsible for replicating the chunks across other data nodes.The default block size in HDP (128 Mb).The default replication factor is 3.
  + Name nodes and Data nodes are components of HDFS service.The name node is an HDFS master component while a data node is an HDFS worker component.The name node and data node are implemented as deamons(background processes) running inside java virtual machine(JVM).
* NAME NODE AND DATA NODE INTERACTION :



* + A need for replication may occur for many reasons :
    - A datanode may become unavailable.
    - A replica could become corrupted.
    - The replication factor of a file may be increased.

Link : <https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>

<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>