Name: Biniam Arefaine

ID: 110972

----

1. [2] **Can you think of a use case of Big Data? Explain it briefly. (Do not repeat the ones from the slides!)**

Solutions:

Fraud protections: Fraud protection is very important for credit card holders. Thanks to big data analytics and machine learning, today's fraud prevention systems are orders of magnitude better at detecting criminal activities and preventing false positives. For example, if a credit card was used to purchase or rent any goods in any other location other than the owners living location, then a customer service agent might call to confirm the card holders is using his/her card and not stolen

Data Warehouse Offload: as data tends to grow, data warehouse tends to be very costly to purchase and run. As more reports and insights are demanded from the BI teams, the data warehouse solutions haven't always been able to provide the desired performance. So, as a result, many enterprises are using an open source big data solution like Hadoop to replace or compliment their data warehouse.

Hadoop based solutions provide much faster performance while reducing the licensing fees and other costs.

Recommendation Engines: This use case has a huge impact on businesses like Amazon, YouTube and Netflix. When you are watching a movie or watching a product to purchase or you have purchased a product before, you probably now take it for granted that the websites will suggest similar items or videos you might be interested in. This is all thanks to the use of big data analytics to analyze historical data.

2. [2] **What are the advantages of using Hadoop and HDFS?**

Solutions:

Storing large files: As the data are getting bigger and larger, Hadoop provides a good solution to store very large files of Terabytes, Petabytes and greater.

It is fast and cost effective: for Businesses which are exploded with data sets, Hadoop provides a cost-effective storage solutions. Unlike traditional DBMS which are extremely cost prohibitive to scale to process massive volumes of data, Hadoop provides a very efficient cost-effective storage solutions.

Horizontal scalability: there is no need to build larger clusters in fact we just keep adding more nodes as the data keeps growing.

Failure Tolerance: This is one of the most important features of using Hadoop. Whenever data is sent to individual node, that node is also replicated to other nodes in the cluster which basically mean we have a copy available for use in case of an even failure.

We do not need to build large clusters; we just keep on adding nodes. As the data keeps on growing, we keep adding nodes.

3. [2] **Explain the term block abstraction in Hadoop and state it's advantages.**

Solutions:

HDFS block size usually is of 64MB-128MB(default 128MB) and unlike other filesystems, a file smaller than the block size does not occupy the complete block size's worth of memory.

The block size is kept so large so that less time is made doing disk seeks as compared to the data transfer rate.

Having a block abstraction for a distributed filesystem brings several benefits:

- A file can be larger than any single disk in the network.
- Making the unit of abstraction a block rather than a file simplifies the storage subsystem. So, the storage subsystem only deals with blocks, simplifying storage management: blocks are a fixed size.
- Furthermore, blocks fit well with replication for providing fault tolerance and availability.

4. [2] **What is the meaning of fault tolerance in HDFS and how is it achieved?**

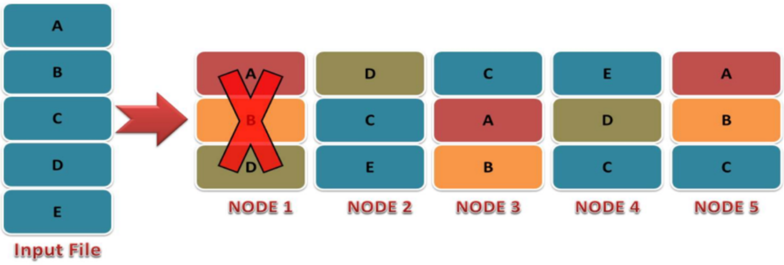
Solutions:

Fault tolerance basically refers to the working strength and capability of a system in unfavorable conditions and how that system can handle such a situation in these particular times.

It is achieved through replication Mechanisms before Hadoop 3 came up. However, Hadoop 3 came up with erasure Coding to achieve Fault tolerance with less storage overhead.

Replication Mechanism:

HDFS creates a replica of the data block and stores them on multiple machines (Data Node).



Erasure Coding:

Erasure coding is a method used for fault tolerance that durably stores data with significant space savings compared to replication.

5. [2] **Consider a 560 TB of text file which needs to be stored in HDFS. The block size has been set to be 128 MB with a replication factor of 3. The cluster has 100 Data Nodes each with a capacity of 15 TB. Will it be possible to store this text file in this HDFS cluster? Why or why not**?

Solutions:

- So, we have a text file of 560TB
- Total capacity will be 15TB * 100 nodes 1500TB.
- And with replication factor of 3 to store 560 TB will be 560 * 3 = 1680TB total space.

Therefore, we cannot store it in this cluster.