Atharva Prashant Pawar (9427) - [Batch D]

BDA : Assignment : 1 DATE: (1)

Q1) Describe the Hadoop Distributed File System (HDFS) & its advantage over traditional file sys.:

=> The Hadoop Distributed File Sys. (HDFS) is a distributed storage sys. designed to store & manage large datasets across clusters of commodity h/w. Unlike traditional file sys., HDFS is optimized for big data applications. It divides large files into smaller blocks (typically 128 MB or 256 MB) & stores multiple copies of these blocks across different nodes in the cluster for fault tolerance.

* Advantages of HDFS over traditional file sys. include:

(A) Scalability:
HDFS scales easily by adding more nodes to the cluster, accommodating the growth of data seamlessly.

(B) Fault Tolerance:
Data redundancy ensures data availability even if individual nodes fail.

(C) High Throughput:
HDFS is optimized for streaming data access, making it ideal for large-scale data processing.

(D) Cost Efficiency:
Utilizing commodity h/w reduces infrastructure costs compared to specialized storage solutions.

(E) Parallel Processing:
HDFS enables distributed data processing, speeding up tasks like map-reduce operations.

(F) Data Locality:
HDFS stores data close to the processing nodes, minimizing data transfer & improving performance.

DATE:

Q.2> What is MapReduce? How does it work in Hadoop? Provide a simple example to illustrate the concept.

=>

MapReduce is a programming model & processing framework used in Hadoop for large-scale data processing. It breaks down complex tasks into smaller, parallelizable tasks & then combines their results to solve problems efficiently. MapReduce operates in two main phases: the Map phase & the Reduce Phase.

* Map Phase:

Input data is divided into smaller chunks & a map function processes each chunk independently, generating intermediate key-value pairs. These intermediate key-value pairs are grouped by their keys & a reduce fun. aggregates & processes the values associated with each key, producing the final O/p.

-eg: Word Count.

* We have a large text doc. & want to count the freq. of each word.

A. Map Phase:

- Mapper splits the text into words: "Hello World, Hello Hadoop"
- Mapper emits key-value pairs: ("Hello", 1), ("world", 1), ("hello", 1), ("Hadoop", 1).

* Shuffle & Sort:

- Framework groups & sorts the intermediate key-value pairs by key: ("Hello", [1, 1]), ("Hadoop", [1]), ("world", [1]).

B. Reduce phase:

- Reduce processes each grp of values: ("Hello", [1, 1]) → ('Hello', 2), ('Hadoop', [1]) → ("Hadoop", 1), ('world', [1]) → ('world', 1).

* Final o/p:
- Word count results : ("Hello", 2), ('Hadoop', 1), ('World', 1).

* MapReduce's parallel processing & fault tolerance make it suitable for big data analysis, as it can efficiently process vast amounts of data across distributed clusters while handling node failures gracefully.

Q3) Write map & reduce steps based on Sales Data Analysis.
=>

A) Map Step for Sales Data Analysis :
For Sales Data Analysis, the i/p data, which consists of sales records, is processed by individual mappers. Each mapper reads a sales record & extracts relevant info., such as the product category & sales amount. The mapper then emits key/value pairs, where the key is the product category & the value is the sales amount associated with that category.

* eg:
i/p : (Date: 2023-08-13, Product: Laptop, Category: Electronics, Sales Amount: Rs 1000)
o/p Mapper o/ps : (key: Electronics, Value: Rs 1000)

B) Reducer step for Sales Data Set Analysis:
The o/p of the map step is processed. The framework groups the emitted key-value pairs by the product category & sends each group to a reducer. The reducer then iterates through the values associated with each category, calculating the total sales amount for that category. The reducer emits the final result, which includes the product category & the total sales amount.

DATE:

* eg:
Mapper o/p: (key: Electronics, Value: [Rs 1000, Rs 800, Rs 1200])
Reducer o/p: (key: Electronics, Total Sales: Rs 3000]

Q4) Facebook Case Study: Social N/w Analysis:

a) How does FB use Hadoop to process & analyze user-generated content & interactions?

⇒ FB employs Hadoop to process & analyze user-generated content & interactions through its data processing framework called "Facebook Analytics for Apache Hadoop" (FAH). FAH leverages Hadoop's distributed computing capabilities to manage & process massive amounts of data. It uses the Hadoop Distributed File System (HDFS) to store data & utilizes tools like MapReduce & Hive for querying & analysis. This enables to gain insights into user behaviours, interests, & trends, enhancing their ability to personalize user experiences & make data driven decisions.

b) What benefits does FB gain from analyzing user behavior & how does it impact user engagement & platform improvements?

⇒ Analyzing user behavior benifits FB in several ways. It helps understand user preferences, content consumption patterns, & engagement levels. This data-driven approach enables FB to tailor content recommendations, improve ad targeting & optimize user engagement. Insights derived from user behavior analysis guide product enhancements & new features developments, ensuring the platform remains relevant & appealing to users. This iterative process fosters user satisfaction & loyalty, ultimately driving increased user engagement & prolonged platform usage.

c) Can you provide examples of specific insights that facebook might have gained through its Hadoop based analysis?

=>

Specific insights FB might have gained include identifying trending topics or hashtags among users, determining the effectiveness of content types (text, images, video) in driving engagement, analyzing the geographic distribution of user activity, detecting patterns of connections & interaction b/w users & assessing the impact of algo. changes on content visibility. For instances, analyzing interactions with posts related to a certain event could help FB understand user sentiment & preferences. Furthermore, analyzing the click-through rates of ads enables them to refine ad targeting strategies. Such insights help Facebook refine its algo., optimize content delivery & enhance user experience.

Q5) Healthcare Case Study: Genomic Data Analysis

a) How is Hadoop utilized in the healthcare industry for analyzing genomic data?

=> Hadoop plays a crucial role in the healthcare industry for analyzing genomic data. Genomic data, which includes info. about an individual's DNA seq. is voluminous & complex. Hadoop's distributed storage (HDFS) & processing capabilities enable healthcare organizations to efficiently store, manage & process massive genomic datasets. Tools like Apache Spark & HBase within the Hadoop ecosystem help analyze genetic variations, identify disease markers & uncover patterns within the data, facilitating medical research & advancements.

(6)

DATE:

b) What challenges does Hadoop help address in managing & analyzing large-scale genomic datasets?

→ Hadoop addresses significant challenges in managing & analyzing large-scale genomic datasets. Genomic data is enormous, making traditional storage & analysis methods inadequate. Hadoop's scalability allows healthcare institutions to store & process petabytes of genomic info. efficiently. Its fault tolerance ensures data integrity, crucial for maintaining accuracy in sensitive medical research. Additionaly Hadoop's parallel processing capabilities speed up analyses that would be prohibitively time consuming with conventional approaches.

c) Discuss the potential impact of Hadoop-based genomic data analysis on personalized medicine & drug discovery.

→ Hadoop-based genomic data analysis has transformative potential for personalized medicine & drug discovery. By analyzing an individual's genetic makeup, doctors can tailor treatment plans based on a patient's genetic predispositions, increasing treatment efficacy & minimizing side effects. Hadoop's data processing helps identify disease-causing mutations & genetic risk factors, enabling early disease detection. In drug discovery, Hadoop allows researches to analyze vast datasets of genomic & pharmaceutical info., leading to the identification of potential drug targets & personalized therapies. This approach accelerates drug development & reduces trial & error in pharmaceutical research, ultimate revolutionizing healthcare outcomes.