CMPE 272 - Apr 25th 2024

Teammates: SJSU ID

Satwik Upadhyayula:017423796

Vineet Samudrala : 017426253

Bala Supriya Vanaparthi:017464135

Shreekar Kolanu : 017406493

Part 1: Evaluating Hadoop-able Problems

a. Implementation of "Products related to this item" on Amazon product page

Explanation:

This feature involves generating recommendations for products that are related to the one a user is currently viewing. For example, if a user is looking at a specific video game, Amazon might suggest similar games or accessories.

Is it a good Hadoop-able problem?

Yes. This is well-suited for Hadoop because:

Large-scale Data Processing: The task involves processing extensive data from user interactions, product information, and purchase histories. Hadoop is designed to handle such large datasets efficiently.

Batch Processing: Recommendations can be updated periodically (e.g., daily or weekly), making it suitable for Hadoop's batch processing capabilities.

Scalability: As Amazon's product catalog and user base grow, Hadoop's distributed system can scale to accommodate increasing data volumes.

b. Implementation of "Similar brands" on an Amazon product page

Explanation:

This feature suggests brands that are similar to the one currently being viewed. For instance, if a user is viewing products from a specific brand, Amazon might recommend other brands that offer similar products.

Is it a good Hadoop-able problem?

Yes. This is a good fit for Hadoop because:

Analysis of Brand Data: It requires analyzing large amounts of data to determine brand similarities, a task that Hadoop can manage effectively.

Data Integration: Hadoop excels at integrating data from different sources, such as product descriptions and customer reviews.

Batch Analysis: Identifying similar brands can be done periodically, aligning well with Hadoop's strength in batch processing.

c. Implementation of "Today's Deals" page on Amazon

Explanation:

This feature highlights current deals and discounts available on Amazon, showcasing time-sensitive offers to users.

Is it a good Hadoop-able problem?

Not ideal. While Hadoop could be used, it is not the best fit because:

Real-time Data Processing: "Today's Deals" requires real-time or near-real-time updates to reflect current discounts, which Hadoop is not optimized for.

Low-latency Requirements: Users expect to see the most up-to-date deals with minimal delay, a requirement better met by real-time processing systems like Apache Kafka or Apache Flink.

Frequent Updates: Deals change frequently throughout the day, necessitating more real-time capabilities than Hadoop's batch processing provides.

Part 2: Good Hadoop-able Problem

Topic: Analyzing User Sentiment from E-commerce Reviews

Detailed Description:

Objective:

Develop a system to analyze and categorize user sentiment from product reviews on an e-commerce platform, providing insights into customer satisfaction and aiding in product improvement.

Steps for Implementation:

1. Data Collection:

   Gather product reviews from the e-commerce platform.

   Store the reviews in Hadoop Distributed File System (HDFS) to leverage its scalability and fault tolerance.

2. Data Preprocessing:

   Use MapReduce to clean the text data by removing stop words, punctuation, and other noise.

   Tokenize and transform the text data into a format suitable for analysis, such as converting words into numerical vectors.

3. Sentiment Analysis:

   Utilize Apache Spark in conjunction with Hadoop to perform sentiment analysis on the

reviews.

Implement machine learning algorithms (e.g., Naive Bayes, Support Vector Machines) using libraries like Apache Mahout or Spark MLlib.

Train the machine learning model on a labeled dataset of reviews with predefined sentiments (e.g., positive, negative, neutral).

4. Categorization:

Use the trained model to categorize the reviews into different sentiment buckets.

Store the categorized results back in HDFS for further analysis and reporting.

5. Analysis and Reporting:

Use Apache Hive or Apache Pig to query and analyze the sentiment data.

Generate reports to identify trends, such as common issues highlighted in negative reviews or features praised in positive reviews.

6. Visualization:

Use Apache Zeppelin or integrate with a BI tool like Tableau to visualize the sentiment analysis results.

Create interactive dashboards to help stakeholders understand customer sentiment and make informed decisions.

Benefits of Using Hadoop:

Scalability: Can handle the vast amount of review data generated by an e-commerce platform.

Fault Tolerance: Ensures data is not lost in case of hardware failures.

Cost-Effectiveness: Utilizes commodity hardware, reducing overall costs.

Integration with Machine Learning: Leverages Hadoop's ecosystem to implement and execute machine learning algorithms efficiently.

Conclusion:

Analyzing user sentiment from e-commerce reviews is an excellent Hadoop-able problem due to the large data volumes involved, the need for batch processing, and Hadoop's robust ecosystem that supports various stages of data processing and analysis.