Big Data Introduction

January 25, 2025

"There is no perfect solution, only trade-offs. Every choice comes with its own costs and compromises." — Thomas Sowell

*A trade-off involves compromising or sacrificing one thing for another.

Outline

Trade Offs

- Distributed Computing vs. Centralized Systems
- Scaling Vertically vs. Horizontally
- Batch vs. Stream Processing
- Monolithic vs. Microservices
- Inter-Service Communication (EDA or REST API)
- Programming Language (Python, Java, Golang, C, C++, etc ..)
- Database (SQL, No-SQL, GraphDB etc ..)
- File Formats
- ...
- ...

Let us start : Binary Search vs Linear Search

Big Data Vs Traditional Data

Discussion on Big Data Vs Traditional Data

{Volume, Velocity ....}

Big Data and DSA Relations - Not Exhaustive

Some ...

- Data Structures
  - HashMap - Dictionary in Python.
  - Trees.
  - Graphs.
- Algorithm Design (Divide and Conquer)
  - Batch Processing
  - Distributed Data Processing
  - Concurrent processing

MapReduce Paradigm uses Divide and Conquer + HashMap

Search Engines: Trie, B-trees for indexing.
Networks: Graphs for Nodes and Connections - Edges and Vertices.
...
...

Batch Processing Systems

- A batch processing system:
  - Takes a large amount of input data.
  - Chunk the large data into batches
  - Runs a job to process the data.
  - Produces output data.
- Jobs often take a while (from a few minutes to several days), so there usually isn't a user waiting for the job to finish.
- Batch jobs are often scheduled to run periodically (e.g., once a day).
- The primary performance measure of a batch job is usually throughput (time to process an input dataset of a certain size).

Other Systems

- Services (online systems) - Request/ Response
- Stream processing systems (near-real-time systems)

Simple Coding Questions - 1

Basic Batch Processing with Lists

How would you implement a batch processing mechanism for a list of numbers

Split the list of numbers into batches
Find the sum of each batches
Combine all the sums and get the final result

Example
data = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
batchSize = 5
Batch [1, 2, 3, 4, 5]: Result = 15
Batch [6, 7, 8, 9, 10]: Result = 40
Batch [11, 12, 13, 14, 15]: Result = 65
Overall Result 120

Solution 1

Solution 1

Some of the key concepts

- List Slicing. – list[start:stop:step]
- Yield vs Return

Simple Coding Questions - 2

Batch Processing of a Large File

You have a large text file. First Part :
Read the file and find out how many vowels in the file.
Second Part :
Read the file - 1000 lines as batch
Find the vowels in each batch
And find out the sum of the vowels from all batches

Solution 2

Solution 2.1 - Read Entire File
Solution 2.2 - Read Files as Chunks
Solution 2.3 - Concurrent Programming
Solution 2.4 - MapReduce With Single Key
Solution 2.5 - MapReduce With Multiple Keys

Some of the key concepts

- File Reading - "With" Keyword
- Batch processing of a file
- Concurrent Programming
- MapReduce Paradigm
- Hashkey / Dictionary - Python

Assignment 1

Merge Sort

- Perform merge sort as usual
- Use Batch Processing we did in the above exercises
- Can you try to attempt MapReduce Paradigm for this ?

Assignment 2

Currency Analysis

- Collect the INR to USD data for the past X years
  - 10, 20 and up to 50 years
  - If possible Use python script or APIs for scraping the values
  - Else get the data by copying from the website and store it in csv
- Analyze the trend - Whether INR is weak or USD is strong ?
  - Device an algorithm to find the peaks
  - Compare with other currencies (Euros vs USD)
  - Figure out whether INR is weak or USD is strong
- Use Batch Processing for the above as we did in the above exercises
- Can you try to attempt MapReduce Paradigm for this ?

Overloads - Large number of Requests

- Rate Limiting ..

Thank You!

Questions?

References

- Designing Data-Intensive Applications by Martin Kleppmann Released March 2017 Publisher(s): O'Reilly Media, Inc. ISBN: 9781491903100
- to be added