ABHIMANYU BHOWMIK

A910119819008 || B.TECH AI

Big Data Home Assignment

Problem Statement: Marks:10

Use Case 1: 25GB date of student's details including general info & academic score

Use Case 2: In the admission year 2023, from May-June, the number of hits to the university website is equivalent to 4.5 GB of data.

Use Case 3: Amity university has self-campus in all the states of India & 10 more best cities in the world. The full record was analysed to invest in the next academic year, only on 1 university campus.

Questions:
Identify the best use-case which suits Big-Date architecture. Mentions support your answer with reasons.

Design the data (multi-variant) in Hadoop with specific details for the number of data nodes, block size, and a few implementation details if possible

Solution:

Identifying Usecase:

The best use case which suits big data architecture is use case 3. Because according to the 3 Vs of big data, i.e.

Volume: High quantity of data.

Velocity: High volume/time of data.

Variety: High variety of data i.e. combination of structured, unstructured, and semistructured data.

Use Case 1 Consist of only volume.

Use Case 2 Consist of high velocity of data.

Use Case 3 is a mix of high volume and variety of data. Though it does not describe what is the frequency of data collection, It can be assumed that such a high volume of all university data will be sent to the main server quite frequently (weekly or monthly).

Designing Big Data Architecture:

To design a viable solution for big data architecture in this particular use case 3, we need to understand what are the data source and data type along with what will we do with the data.

If we what to understand the sources of production of data, it is very much clear in the problem statement itself. i.e. The university campuses located in,

1. India (28)
2. Outside of India (10)

We should have Multiple data warehouses inside of India and at least 1 outside of India to process university data abroad.

A viable solution is 1 data warehouse for every 10 campuses.

The Proposed solution for Data warehouses:

a. Northern Indian (North and West India)
b. East Indian (East - Central India)
c. South Indian (All the South Indian States)
d. Abroad (All 10 universities outside India)

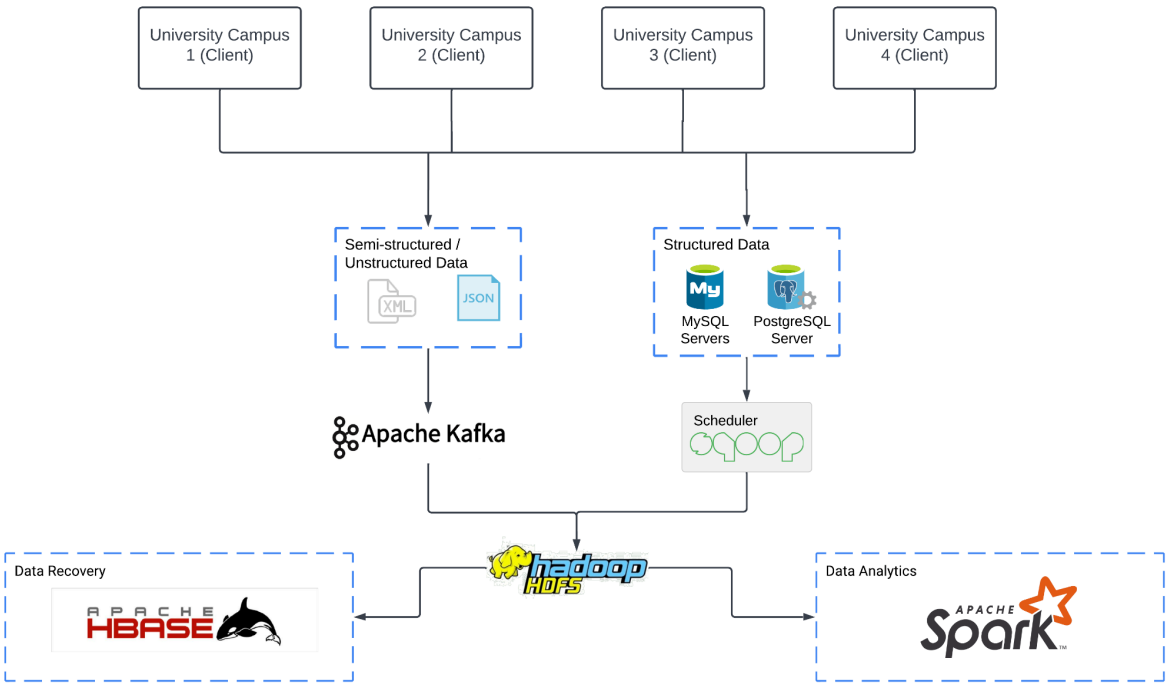
From the problem statement, it is not very clear which kind of data will generate from universities. Though we can assume some categories of data which can be generated. E.g.

- Structured Data (Student, Faculties, staff and Academic, Operational, Accounts Details etc.) -> SQL
- Documents Data (Notice, Legal documents, Personal documents of students and faculty, Teaching Documents, Examination papers etc.) -> PDF, Word
- Semi-structured Data (University Social engagement Data, Library, canteen and student engagement data etc.) -> JSON, XML
- Text Data (Emails, Internship opportunities, On-campus Placement Data etc.)
- Audio - Video Data (Details of Fest, New campus, buildings, Infrastructure Inspection Data etc.) -> MP3, MP4

Among all these data structured and semi-structured data are very useful for analysing a particular university on the basis of various parameters. For these purposes, we can choose the structure and semi-structured data to store in Big-Data Architecture for analytical processing.

The Proposed Big-Data Architecture:

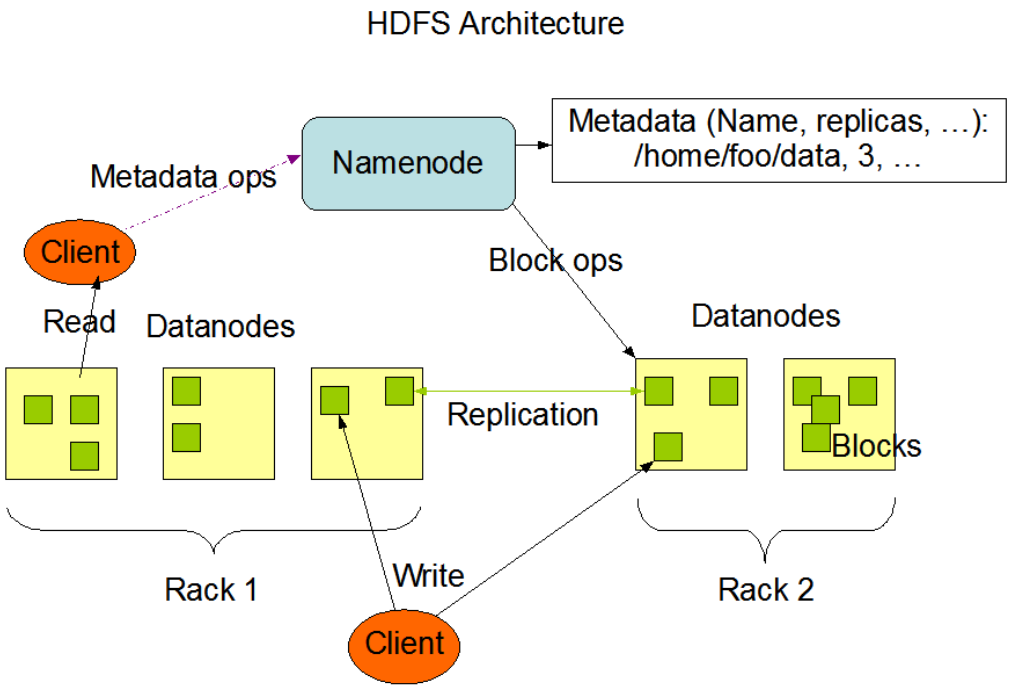
1. Structured Data -> Apache Sqoop Schedular -> HDFS -> Apache Spark (Analytics)
2. Semi/Un - Structured Data -> Apache kafka -> HDFS -> Apache Spark(Analytics)
3. Apache Hbase -> For custom Data Recovery



The proposed method should be implemented in each and every Data warehouse.

For particular HDFS implementation, we can follow some estimations. (As data volume is not mentioned and any specific criteria we can't specifically determine the block size and No. of datandes and number of replicas). In this case, we assume a max of 500 TB (specifically 512 TB) of data storage for each University Campus.

The standard HDFS architecture is given below:



Specific Implementation Details:

Block Size: 512 MB

Replication Factor: 3

Name-Node: 2 Namenode for Each 5 Client
1. Active Namenode
2. Standby Namenode (Fault tolerance)
Active + Standby namenode works as a single entity
[Total 4 name nodes per Data warehouse]

No. Blocks in Datanodes: 3000

Datanode: Each Namenode can handle up to 5000 Datanodes, so each data warehouse can have up to 10k data nodes.

Total Storage for Each Client: (Block Size x No. Blocks x Max. Datanodes) / (Rep. Factor x No. Clients)

= (512 x 3000 x 5000) / (3 x 5)
= 51,20,00,000 MB
= 512 TB