University of East London

Pioneering Futures Since 1898

Tutorial 1: Big Data Analytics - CN7031

Big Data Processing with Hadoop
Case Study: RetailSphere

Learning Outcomes

- Understand the 5Vs of big data and their implications.
- Explain the functionality and advantages of HDFS and YARN.
- Analyze the MapReduce process and its optimization.
- Evaluate the importance of data formatting and partitioning.
- Address legal, social, ethical, and professional issues in data management.

Dr Amin Karami
Associate Professor in AI & Data Science
CDT School, Dockland Campus, EB.1.98

email: a.karami@uel.ac.uk
web: www.aminkarami.com

October 2, 2025

Case Study: RetailSphere

RetailSphere, a leading global retailer, processes millions of transactions daily across various channels. They aim to analyze customer purchase patterns to optimize stock levels, personalize marketing, and improve overall customer satisfaction. The challenge is to efficiently manage the vast volume, variety, and velocity of data while ensuring data veracity and extracting value. They are using Hadoop for storage and processing. **Please complete the multiple choice questions by considering this case study.**

Multiple Choice Questions

1. RetailSphere's need to process transactions in real-time exemplifies which 'V' of big data?
   (a) Volume
   (b) Variety
   (c) **Velocity**
   (d) Veracity
2. Ensuring the accuracy and trustworthiness of RetailSphere's data relates to which 'V'?
   (a) Volume
   (b) Variety
   (c) Velocity
   (d) **Veracity**
3. Why is data formatting crucial before storing data in Hadoop?
   (a) **To ensure compatibility and efficient processing**
   (b) To improve data security
   (c) To reduce storage costs
   (d) To increase data velocity
4. What is a major advantage of using HDFS for RetailSphere?
   (a) Centralized data storage
   (b) **High fault tolerance and scalability**
   (c) In-memory data processing
   (d) Real-time analytics
5. How does HDFS ensure data reliability for RetailSphere's vast datasets?
   (a) By using a single server for all data
   (b) **By replicating data across multiple nodes**
   (c) By compressing data
   (d) By encrypting data
6. What is the primary function of the Mapper in MapReduce for RetailSphere's data processing?

(a) Aggregate data
(b) **Extract key-value pairs**
(c) Visualize data
(d) Encrypt data

7. What critical role does the Reducer play in RetailSphere's MapReduce jobs?
   (a) To split data into smaller chunks
   (b) **To summarize and aggregate data**
   (c) To store data in HDFS
   (d) To shuffle data
8. Why is partitioning crucial for RetailSphere's Hadoop jobs?
   (a) It enhances data security
   (b) It reduces data volume
   (c) **It facilitates parallel processing**
   (d) It simplifies data visualization
9. How does partitioning contribute to improved performance in Hadoop?
   (a) By encrypting data
   (b) **By allowing simultaneous data processing**
   (c) By storing data centrally
   (d) By compressing data
10. The shuffle and sort phase in MapReduce is crucial because:
    (a) It enhances data security
    (b) **It organizes data for efficient reduction**
    (c) It compresses data
    (d) It visualizes data
11. What is a common issue during the shuffle phase that RetailSphere needs to address?
    (a) Data encryption errors
    (b) **Network bottlenecks and latency**
    (c) Inaccurate data sorting
    (d) Excessive data replication
12. Which strategy can RetailSphere use to optimize their Hadoop processes?
    (a) Increasing data replication
    (b) Reducing the number of data nodes
    (c) **Optimizing MapReduce job parameters**
    (d) Using a single data server
13. To improve job performance, RetailSphere should focus on optimizing:

(a) **The number of Mappers and Reducers**

(b) The amount of data stored

(c) The encryption level of data

(d) The visualization tools used

14. Which ethical consideration is crucial for RetailSphere when handling customer data in Hadoop?

(a) Data encryption

(b) Data accuracy

(c) **Data privacy and consent**

(d) Data volume

15. What legal issue must RetailSphere consider when storing customer data internationally?

(a) Data sorting

(b) Data replication

(c) **Compliance with data protection laws**

(d) Data visualization

16. What is the primary function of YARN in the Hadoop ecosystem?

(a) It manages data storage

(b) It provides real-time analytics

(c) **It manages resources and job scheduling**

(d) It enhances data encryption

17. Which component of YARN is responsible for monitoring the status of resources and nodes?

(a) ResourceManager

(b) **NodeManager**

(c) ApplicationMaster

(d) DataNode

18. What is the purpose of the Intra-DataNode Balancer in HDFS?

(a) To encrypt data within nodes

(b) **To balance data load within a DataNode**

(c) To replicate data across nodes

(d) To compress data for storage

19. Why is data replication important in Hadoop?

(a) To enhance data processing speed

(b) **To improve fault tolerance and availability**

(c) To reduce storage costs

(d) To simplify data encryption

20. How does proper data distribution benefit RetailSphere's Hadoop operations?

    (a) By enhancing data visualization
    (b) **By enabling even data load across the cluster**
    (c) By reducing data replication needs
    (d) By simplifying data compression

21. Increasing the block size in HDFS can lead to:

    (a) Increased data redundancy
    (b) Reduced network congestion during shuffling
    (c) **Improved MapReduce efficiency by reducing seeks**
    (d) Decreased fault tolerance

22. RetailSphere is considering using Hadoop's YARN. What is a key benefit of YARN for their operations?

    (a) It provides real-time data analytics
    (b) **It separates resource management from processing**
    (c) It simplifies HDFS replication
    (d) It enhances data encryption

23. What is a potential downside of excessive data replication in HDFS for RetailSphere?

    (a) Increased data processing speed
    (b) **Higher storage costs**
    (c) Improved fault tolerance
    (d) Reduced network latency

24. Which of the following is an advantage of using parallel processing in Hadoop for RetailSphere?

    (a) Increased data redundancy
    (b) **Faster data processing**
    (c) Simplified data formats
    (d) Reduced need for data encryption

25. How can RetailSphere ensure data veracity in their Hadoop ecosystem?

    (a) By optimizing data visualization
    (b) **By implementing data validation checks**
    (c) By increasing data volume
    (d) By using advanced encryption

26. In Hadoop, what is the role of the NodeManager in YARN?

    (a) To manage data storage

(b) **To monitor and report on node resource usage**

(c) To encrypt data at rest

(d) To visualize data processing

27. Why is network bandwidth a critical consideration during the shuffle phase in MapReduce?

(a) It affects data visualization speeds

(b) **It can become a bottleneck, slowing down data processing**

(c) It reduces data redundancy

(d) It enhances data encryption

28. Which component in YARN is responsible for negotiating resources from the ResourceManager?

(a) DataNode

(b) NodeManager

(c) **ApplicationMaster**

(d) NameNode

29. How does the Intra-DataNode Balancer improve Hadoop performance for RetailSphere?

(a) By encrypting data within DataNodes

(b) **By balancing storage load within a DataNode**

(c) By increasing data block size

(d) By reducing data replication needs

30. What is an essential consideration for RetailSphere when optimizing parallel processing in Hadoop?

(a) Increasing data block size

(b) **Balancing the load across all nodes**

(c) Reducing the number of Mappers

(d) Using a single server for processing