


Name:			
Enrolment No:			
UNIVERSITY OF PETROLEUM AND ENERGY STUDIES End Semester Examination, May 2022			
Course: Big Data Search Program: B. Tech, (CSE) SPZ Big Data Course Code: CSEBD3004		Semester: VI Time: 03 hrs. Max. Marks: 100	
Instructions: Attempt all questions.			
SECTION A (5Qx4M=20Marks)			
S. No.	Statement of question	Marks	CO
Q 1	What makes data, “Big” Data? Why is Big data important?	4	CO1
Q 2	How does Solr Highlighting work? Give an example.	4	CO2
Q 3	What does caching mean in the context of spark streaming?	4	CO3
Q 4	How do you prevent split-brain problems in the Elasticsearch cluster?	4	CO4
Q 5	How would you describe Lucene's core query processing classes?	4	CO5
SECTION B (4Qx10M= 40 Marks)			
Q 6	How would you define Apache Solr? Describe Apache Solr's internal architecture.	2+8	CO2
Q 7	Explain Kafka’s messaging model. How do records flow in Apache Kafka?	5+5	CO3
Q 8	How would you define an Index and a document in Elasticsearch? Why are shards used in Elasticsearch?	5+5	CO4
Q 9	How would you define Lucene? How does Lucene build an index? Give an example.	5+5	CO5
OR			
	How do you define the roles of analyzers in the Lucene search engine? Describe the different types of analyzers available in Apache Lucene.	5+5	CO5

SECTION-C
(2Qx20M=40 Marks)

Q10	<p>(a) What are the major issues involved in search engine design? Explain major components of the building blocks of a search engine.</p> <p>(b) Draw the inverted index that would be built for the following document collection.</p> <p style="text-align: center;">Doc 1 new home sales top forecasts Doc 2 home sales rise in July Doc 3 increase in home sales in July Doc 4 July new home sales rise</p>	10+10	CO1
Q11	<p>(a) How would you explain the working of Spark architecture? Write all steps.</p> <p>(b) Define Resilient Distributed Datasets (RDD). How would you describe the RDD's operations?</p> <p style="text-align: center;">OR</p> <p>(a) Write main role of Streaming Context in Spark. Why need to apply transformations on DStream?</p> <p>(b) How do you explain importance of accumulators, broadcast variables and caching during Spark streaming?</p>	10+10	CO3