UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2019

Course: Big Data Ingestion
Program: BTech CSE + Big Data
Course Code: CSBD 2002

Semester: IV
Time 03 hrs.
Max. Marks: 100

Instructions: Attempt all questions from section A. There is an internal choice in question 9 and 11

SECTION A

S. No.		Marks	CO
Q1	Distinguish between Sqoop1 and Sqoop2?	4	CO4
Q2	Compare between --target-dir and --warehouse-dir arguments? Can we use both in the same import command? Use of the warehouse directory in Sqoop import?	4	CO4
Q3	Import an RDBMS table **order_details** (in an **order_db** MySQL database) in HDFS using Sqoop when we do not have a primary key column.	4	CO4
Q4	a) Summarize the importance of using the split-by clause in Apache Sqoop? b) What is the significance of the **eval** tool?	4	CO4
Q5	Explain the configuration parameters i. batch.size ii. linger.ms iii. max.in.flight.requests.per.connection iv. max.partition.fetch.bytes	4	CO3

SECTION B

Q6	Define compaction in Kafka and how does it work?	10	CO4
Q7	Design and illustrate the Big Data Ingestion Architecture. What are the different Big Data Ingestion challenges?	10	CO1
Q8	Explain a. Controller b. Replication c. Partition Allocation d. Compaction	10	CO2
Q9	a) Categorize the three different partition assignment strategies from a consumer	10	CO3

	perspective in a consumer group? **OR** b) Describe Unclean Leader Election policy and Minimum In-Sync replicas		
	SECTION-C		
Q10	a) There are 300 tables in a database name **order_db**. We want to import all the tables from the database except the tables named Table130, Table140, and Table150. How can we do this without having to import the tables one by one? (6) **b)** Import all records from the **customer** table (in a **retail_db** MySQL database) in HDFS. The data format of transferred data should be suitable for BI Tools. It should also be compressed in Snappy compression codec. (6) c) Import data from **items** table (in an **order_db MySQL** database) where the category is between 3 and 20 and data should be stored in **item_subset** in HDFS folder. (6) d) Discuss how does the Kafka preserve the order of messages within a partition? (2)	20	CO1, CO4
Q11	a) Construct the Kafka Architecture in detail. What is the role of the Zookeeper in the Kafka cluster? **OR** b) Sketch the Kafka Producer components diagram and illustrate the different steps involved in the sending the data to Kafka broker.	20	CO2, CO3