Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

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

Course: Big Data Analytics — **Semester: VIII**
Program: B.Tech. CS-OGI — **Time 03 hrs.**
Course Code: CSIB385 — **Max. Marks: 100**
Instructions: **All sections are compulsory** — **Nos. of page(s) : 3**

SECTION A

S. No.	Write short notes on the following	Marks	CO
Q 1	Identify the 5 V's of Big Data.	4	CO1
Q 2	Differentiate between Semi structured and Unstructured data	4	CO2
Q 3	Identify the different components of Hadoop ecosystem	4	CO3
Q 4	Explain the concept of data lake.	4	CO3
Q 5	What is federated data?	4	CO4

SECTION B

	All questions are compulsory	Marks	CO
Q 6	Explain the big data framework with diagram	10	CO4
Q 7	Differentiate between RDBMS & Hadoop	10	CO1
Q 8	Explain the Map Reduce Isolation process	10	CO3
Q 9	Explain in detail the working of HDFS **OR** Identify the key configuration of HDFS	10	CO4

SECTION-C

	Case Study	Marks	CO
	Yelp was founded in 2004 with the main goal of helping people connect with great local businesses. The Yelp community is best known for sharing in-depth reviews and insights on local businesses of every sort. In their ten years of operation Yelp went from a one-city wonder (San Francisco) to an international phenomenon spanning 29 countries and more than 120 markets. As of June 2014, Yelp had an average of 138 million monthly unique visitors and more than 61 million local reviews have been written by yelpers.		

The Challenge

Yelp has established a loyal consumer following, due in large part to the fact that they are vigilant in protecting the user from shill or suspect content. Yelp uses an automated review filter to identify suspicious content and minimize exposure to the consumer. The site also features a wide range of other features that help people discover new businesses (lists, special offers, and events), and communicate with each other. Additionally, business owners and managers are able to set up free accounts to post special offers, upload photos, and message customers.

The company has also been focused on developing mobile apps and was recently voted into the iTunes Apps Hall of Fame. Yelp apps are also available for Android, Blackberry, Windows 7, Palm Pre and WAP.

Local search advertising makes up the majority of Yelp's revenue stream. The search ads are colored light orange and clearly labeled "Sponsored Results." Paying advertisers are not allowed to change or re-order their reviews.

Why Amazon Web Services

Yelp originally depended upon giant RAIDs to store their logs, along with a single local instance of Hadoop. When Yelp made the move to Amazon Elastic MapReduce (Amazon EMR), they replaced the RAIDs with Amazon Simple Storage Service (Amazon S3) and immediately transferred all Hadoop jobs to Amazon Elastic MapReduce.

"We were running out of hard drive space and capacity on our Hadoop cluster," says Yelp search and data-mining engineer Dave Marin.

Yelp uses Amazon S3 to store daily logs and photos, generating around 1.2TB of logs per day. The company also uses Amazon EMR to power approximately 20 separate batch scripts, most of those processing the logs. Features powered by Amazon Elastic MapReduce include:

- People Who Viewed this Also Viewed
- Review highlights
- Auto complete as you type on search
- Search spelling suggestions
- Top searches
- Ads

Their jobs are written exclusively in Python, while Yelp uses their own open-source library, mrjob, to run their Hadoop streaming jobs on Amazon EMR, with boto to talk to Amazon S3. Yelp also uses s3cmd and the Ruby Elastic MapReduce utility for monitoring.

Yelp developers advise others working with AWS to use the boto API as well as mrjob to ensure full utilization of Amazon Elastic MapReduce job flows. Yelp runs approximately 250 Amazon Elastic MapReduce jobs per day, processing 30TB of data and is grateful for AWS Support that helped with their Hadoop application development.

	The Benefits Using Amazon Elastic MapReduce Yelp was able to save $55,000 in upfront hardware costs and get up and running in a matter of days not months. However, most important to Yelp is the opportunity cost. "With AWS, our developers can now do things they couldn't before," says Marin. "Our systems team can focus their energies on other challenges."		
Q 10	Compare the different other technologies which could have been used in place of Hadoop and benefits of it, if any.	**20**	**CO5**
Q 11	Critically analyze the technologies used in the above scenario. OR Design a Big Data Architecture of your own for solving the above scenario.	**20**	**CO5**

Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

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

Course: Big Data Analytics — **Semester: VIII**
Program: B.Tech. CS-OGI — **Time 03 hrs.**
Course Code: CSIB385 — **Max. Marks: 100**
Instructions: **All sections are compulsory** — **Nos. of page(s) : 3**

SECTION A

S. No.	Write short notes on the following	Marks	CO
Q 1	Explain Flat scalability of Hadoop	4	CO1
Q 2	Identify the different types of data	4	CO2
Q 3	Identify the common drawbacks of RDBMS systems	4	CO3
Q 4	Explain the concept of NOSQL	4	CO3
Q 5	Explain Data Lake.	4	CO4

SECTION B

	All questions are compulsory	Marks	CO
Q 6	Explain the Cluster configuration of Hadoop	10	CO4
Q 7	Identify the benefits of Hadoop over contemporary technologies	10	CO3
Q 8	Describe the map reduce process in detail	10	CO3
Q 9	Identify the working of HDFS **OR** Explain the architecture of HDFS	10	CO5

SECTION-C

	Case Study	Marks	CO
	The International Centre for Radio Astronomy Research (ICRAR) began in 2009 as a joint venture between Curtin University and The University of Western Australia. Based in Perth, Western Australia, ICRAR's 110 employees are currently part of an international effort to develop the biggest radio telescope in the world, known as the Square Kilometre Array (SKA). During its 50-plus year lifetime, the SKA will expand our understanding of the Universe **The Challenge** Once operational, the SKA is expected to gather and process as much data from the sky every day as the world currently produces in a year. The SKA will use this data to make maps of the sky that scientists can use to study the Universe. A single SKA		

	image could be as big as 600 TB, and each sky map will need thousands of images. "We need to address computing challenges that are immeasurable," says Kevin Vinsen, Research Associate Professor at ICRAR. "When it's fully operational in the next decade, depending on the science case, the SKA might collect between 500 TB and 1 PB of imaging data every day. The sheer amount of raw compute power that we need to do that is mind-boggling." To amass compute resources for a series of preliminary experiments, ICRAR formed a community computing initiative called theSkyNet. This initiative allows ICRAR to use spare CPU cycles volunteered by the public to simulate a supercomputer. Vinsen and his colleagues then use the compute power generated by theSkyNet to analyze images of galaxies from the Pan-STARRS1 telescope in Hawaii as part of theSkyNet project. Crowd-sourced computing projects often run into problems matching physical server capacity to the load of incoming data. ICRAR needed to run experiments using theSkyNet in a cost-effective and flexible way that would allow Vinsen and his team to obtain results quickly.		
Q 10	Compare the different other technologies which could have been used in place of Hadoop and benefits of it, if any.	**20**	**CO5**
Q 11	Critically analyze the technologies used in the above scenario. OR Design a Big Data Architecture of your own for solving the above scenario.	**20**	**CO5**