Data Analytics (IT – 3006)
Practice Questions (Unit 1)

Q1. Why traditional databases cannot store big data?

Q2. A finance company wants to evaluate their users, on the basis of loans they have taken. They have hired you to find the number of cases per location and categorize the count with respect to the reason for taking a loan. Next, they have also tasked you to display their average risk score. Discuss and then model your views concerning descriptive and predictive analytics.

Q3. A retail company wants to enhance their customer experience by analysing the customer reviews for different products, so that they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, type of product, etc. You also have to figure out the complaints which have no timely response. Discuss and then model your views concerning descriptive, diagnostic and predictive analytics.

Q4. A mobile health organisation captures patient's physical activities by attaching various sensors on different body parts. These sensors measure the motion of diverse body parts like the acc., the rate of turn, the magnetic field orientation etc. A model will be built for effectively deriving information about the motion of different body parts like chest, ankle etc. Discuss and then model your views concerning diagnostic and predictive analytics.

Q5. Discuss some effective ways to measure the fault tolerance technique to increase the reliability of a private, public and hybrid cloud.

Q6. A new company in Media and Entertainment domain wants to outsource movie ratings & reviews. They want to know the frequent users who is giving review and rating consistently for most of the movies. You have been hired to report and analyze different users, based on which user has rated the most number of movies, their occupations & their age-group by revealing the right questions and seeking answers.

Q7. Design a stock performance system to predict good and bad stocks based on their history exhibiting parallel and distributed computing. Discuss 5 V's in connection to the given problem.

Q8. A social media marketing company wants to expand its business. They want to find the websites which have a low rank web page. You have been tasked to find the low-rated links based on the user comments, likes etc by drawing conceptual data model, logical data model, physical data model and view schema. Also, identify the structured, unstructured and semi-structured data.

Q9. Explain the similarity and difference between JSON and BSON with suitable examples.

Q10. Discuss the challenges associated with semi-structured data and how to deal with it?

Q11. Explain data locality with suitable examples. Explain the difference between "moving computation" and "moving data" in a cluster.

Q12. How distributed and parallel computing play a role in big data environment?

Q13. Outline the main tasks and activities to be performed for each stage in data analytics life cycle for the followings:

- A small stock trading organization, wants to build a Stock Performance System. You have been tasked to create a data model to predict good and bad stocks based on their history. You also have to build a customized product to handle complex queries such as calculating the covariance between the stocks for each month.
- A mobile health organization captures patient's physical activities, by attaching various sensors on different body parts. These sensors measure the motion of diverse body parts like acceleration, the rate of turn, magnetic field orientation, etc. You have to build data model for effectively deriving information about the motion of different body parts like chest, ankle, etc.
- A retail company wants to enhance their customer experience by analysing the customer reviews for different products. So that, they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, type of product, etc. You also have to figure out the complaints which have no timely response.
- A new company in the travel domain wants to start their business efficiently, i.e. high profit for low TCO. They want to analyse & find the most frequent & popular tourism destinations for their business. You have been tasked to analyse top tourism destinations that people frequently travel & top locations from where most of the tourism trips start. They also want you to analyze & find the destinations with costly tourism packages.
- A new airline company wants to start their business efficiently. They are trying to figure out the possible market and their competitors. You have been tasked to analyse & find the most active airports with maximum number of flyers. You also have to analyse the most popular sources & destinations, with the airline companies operating between them.
- A finance company wants to evaluate their users, on the basis of loans they have taken. They have hired you to find the number of cases per location and categorize the count with respect to the reason for taking a loan. Next, they have also tasked you to display their average risk score.
- A new company in Media and Entertainment domain wants to outsource movie ratings & reviews. They want to know the frequent users who is giving review and rating consistently for most of the movies. You have to analyze different users, based on which user has rated the most number of movies, their occupations & their age-group.
- Analyze the Aadhaar card data set against different research queries for example total number of Aadhaar cards approved by state, rejected by state, total number of Aadhaar card applicants by gender and total number of Aadhaar card applicants by age type with visual depiction.
- A salesperson may manage many other salespeople. A salesperson is managed by only one salespeople. A salesperson can be an agent for many customers. A customer is managed by one salespeople. A customer can place many orders. An order can be placed by one customer. An order lists many inventory items. An inventory item may be listed on many orders. An inventory item is assembled from many parts. A part may be assembled into many inventory items. Many employees assemble an inventory item from many parts. A supplier supplies many parts. A part may be supplied by many suppliers.
- A manufacturing company produces products. The following product information is stored: product name, product ID and quantity on hand. These products are made up of many components. Each component can be supplied by one or more suppliers. The following component information is kept: component ID, name, description, suppliers who supply them, and products in which they are used.

Assumptions:

- A supplier can exist without providing components.
- A component does not have to be associated with a supplier.
- A component does not have to be associated with a product. Not all components are used in products.
- A product cannot exist without components.

Q14. Consider the dataset presented below. Draw the MapReduce process to find the maximum electrical consumption for each month.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

Figure 1: Dataset representing electrical consumption for each year

Q15. Describe the working of MapReduce model. Perform the Map Reduce task for the following input files containing the following data.

Input File 1	Input File 2	Input File 3
Apple Orange Mango Orange Grapes Plum	Apple Plum Mango Apple Apple Plum	Apple Orange Mango Plum Apple Grapes

Q16. Consider the following sample data. Draw the MapReduce process to find the number of customers from each city.

customer_id	first_name	last_name	phone	email	street	city	state	zip_code
1	Debra	Burks	NULL	debra.burks@yahoo.com	9273 Thome Ave.	Orchard Park	NY	14127
2	Kasha	Todd	NULL	kasha.todd@yahoo.com	910 Vine Street	Campbell	CA	95008
3	Tameka	Fisher	NULL	tameka.fisher@aol.com	769C Honey Creek St.	Redondo Beach	CA	90278
4	Daryl	Spence	NULL	daryl.spence@aol.com	988 Pearl Lane	Uniondale	NY	11553
5	Charolette	Rice	(916) 381-6003	charolette.rice@msn.com	107 River Dr.	Sacramento	CA	95820
6	Lyndsey	Bean	NULL	lyndsey.bean@hotmail.com	769 West Road	Fairport	NY	14450
7	Latasha	Hays	(716) 986-3359	latasha.hays@hotmail.com	7014 Manor Station Rd.	Buffalo	NY	14215
8	Jacquline	Duncan	NULL	jacquline.duncan@yahoo.com	15 Brown St.	Jackson Heights	NY	11372
9	Genoveva	Baldwin	NULL	genoveva.baldwin@msn.com	8550 Spruce Drive	Port Washington	NY	11050
10	Pamelia	Newman	NULL	pamelia.newman@gmail.com	476 Chestnut Ave.	Monroe	NY	10950

Q17. Consider the following sample data. Draw the MapReduce process to find the number of employees from each category of marital status.

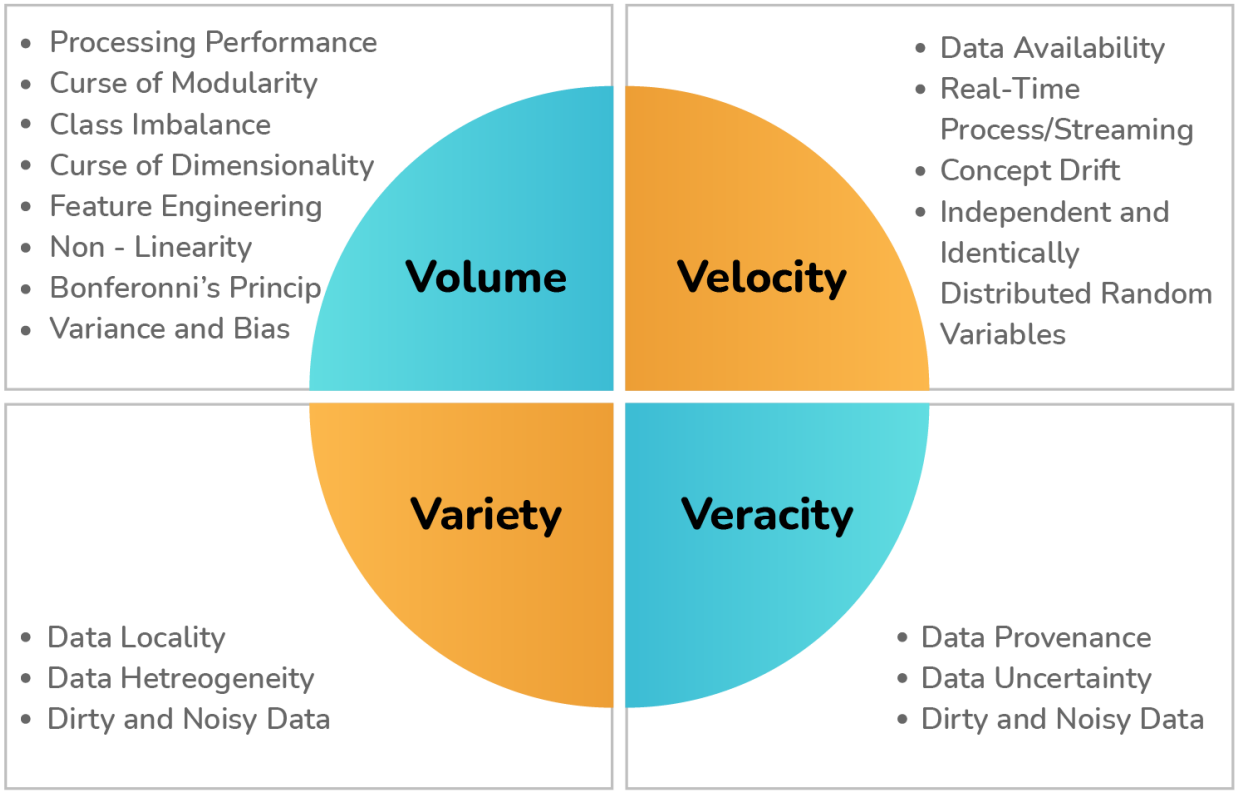
	NationalIDNumber	JobTitle	BirthDate	MaritalStatus	Gender	ModifiedDate
1	295847284	Chief Executive Officer	1969-01-29	S	M	2014-06-30 00:00:00.000
2	245797967	Vice President of Engineering	1971-08-01	S	F	2014-06-30 00:00:00.000
3	509647174	Engineering Manager	1974-11-12	M	M	2014-06-30 00:00:00.000
4	112457891	Senior Tool Designer	1974-12-23	S	M	2014-06-30 00:00:00.000
5	695256908	Design Engineer	1952-09-27	M	F	2014-06-30 00:00:00.000
6	998320692	Design Engineer	1959-03-11	M	M	2014-06-30 00:00:00.000
7	134969118	Research and Development Manager	1987-02-24	M	M	2014-06-30 00:00:00.000
8	811994146	Research and Development Engineer	1986-06-05	S	F	2014-06-30 00:00:00.000
9	658797903	Research and Development Engineer	1979-01-21	M	F	2014-06-30 00:00:00.000
10	879342154	Research and Development Manager	1984-11-30	M	M	2014-06-30 00:00:00.000
11	974026903	Senior Tool Designer	1978-01-17	S	M	2014-06-30 00:00:00.000
12	480168528	Tool Designer	1959-07-29	M	M	2014-06-30 00:00:00.000

Q18. Discuss Hadoop ecosystem by outlining each component.

Q19. Draw a diagram illustrating multi-threaded parallel distributed system.

Q20. Discuss similarities and differences between ELT and ETL.

Q21. Complete the following diagram with the incorporation of value.



Q22. You are planning the marketing strategy for a new product in your business. Identify and list some limitations of structured data related to this work.

Q23. In what ways does analyzing Big Data help organizations prevent fraud?

Q24. Design considerations for distributed systems are: No global clock, Geographical distribution, No shared memory, Independence and heterogeneity, Fail-over mechanism, and Security concerns. Explain each of the terms.

Q25. The reasons to why a system should be built distributed, not just parallel with the characteristics of Scalability, Reliability, Data sharing, Resources sharing, Heterogeneity and modularity, Geographic construction, and Economic. Explain each of the terms in details.

*** The End ***