IST722: Unit 10 Participation Questions

This is an individual assignment.

Before you begin, please make sure you've read and understand 1) our class honor code, 2) course policies on late work and 3) participation policies as posted on the syllabus. "I didn't know" is not an excuse.

You should cite your sources in a standard format like MPA or APA and include a list of works cited.

Your Name:	Akshay Bhala
Your Email:	abhala@syr.edu

Instructions

Answer each of the following questions as concisely as possible. More is not necessarily better. Please justify your answer by citing your sources from the assigned readings from our textbooks, our class lectures, or online if directed to do so. Be sure to cite in text and include a list of works cited. Place your answer below each question. When you're finished, print out this document and bring it to class as part of your participation grade.

Questions

[1] Discuss the rationality behind more data for data-driven decision making.

Ans. More data gives you more insights to curb anomalies and make more accurate decisions. Data-driven decision making is the practice where data is collected, analysed, and decisions are made based on the insights which are derived from the collected information. The process is more objective and can be quickly evaluated according to the influence of the data on metrics. Data-driven decision management is crucial for every industry. It helps the management to plan to see what will speed the production to save time. Data based decision also helps to use past information to predict what is to happen in the future. Without data, there are a lot of risks, such as performing on false assumptions and being swayed by biases.

[2] Explain CAP Theorem of Distributed Systems. Show why it is applicable.

Ans. CAP theorem or Eric Brewers theorem states that we can only achieve at most two out of three guarantees for a database: Consistency, Availability and Partition Tolerance.

Here Consistency means that all nodes in the network see the same data at the same time.

Availability is a guarantee that every request receives a response about whether it was successful or failed. However it does not guarantee that a read request returns the most recent write. The more number of users a system can cater to better is the availability.

Partition Tolerance is a guarantee that the system continues to operate despite arbitrary message loss or failure of part of the system. In other words, even if there is a network outage in the data center and some of the computers are unreachable, still the system continues to perform.

Out of these three guarantees, no system can provide more than 2 guarantees. Since in the case of a distributed systems, the partitioning of the network is must, the tradeoff is always between consistency and availability.

[3] Examine “Schema on Read?” wrt Relational Systems and Big Data Systems.

Ans. In very rapid initial data loading, Schema-on-Read benefits because the data does not have to adopt any internal schema(internal database format) to read or decode or serialize, since it is only a file copy / move. In the situation of massive data or having two schemas for the same underlying data, this method of movement of data is more stable. Flexibility of intent and query power are the key benefits of Schema on Read.

[4] Define Big Data in terms of the 3Vs. Search the internet for 5Vs, 10Vs, 30Vs – what’s the max number you got?

1. VOLUME: Volume, for example, refers to the amount of data generated by websites, portals and online applications within the social media space. Volume covers the available data , particularly for B2C businesses, who are out there and need to be analyzed for relevance. Remember this: Facebook has 2 billion users, 1 billion users on Youtube, 350 million users on Twitter and 700 million users on Instagram. These users add to billions of photos, blogs , videos, tweets and so on every day.

2. VARIETY: Variety in Big Data refers to all the structured and unstructured knowledge that can be generated by humans or machines. Structured documents, tweets, photographs & videos are the most frequently added information. However, under Variation, unstructured material such as texts, voicemails, hand-written text, ECG reading, audio recordings etc. are also important items. The ability to group the incoming data into different categories is all about diversity.

3. VELOCITY: With Velocity, we refer to the rate at which information is produced. In our social media example, 900 million images are uploaded to Facebook every day, 500 million tweets are posted to Twitter, 0.4 million hours of video are uploaded to Youtube and 3.5 billion Google searches are conducted. Big Technology allows the organization to hold this explosion, embrace the incoming data influx and process it rapidly at the same time so that it does not cause bottlenecks.

The other V’s are as follows

Veracity: The fourth V is veracity, which is equal to quality in this sense. We've got all the details, but might there be anything missing? Is that "clean" and reliable data?

Value: Finally, the V for value sits at the top of the big data pyramid. This refers to the ability to transform a tsunami of data into business.

Max number of V’s I got is 42.

[5] Research 3 major differences between Pig and Hive.

	PIG	HIVE
1.	Pig operates on the client side of a cluster.	Hive operates on the server side of a cluster.
2.	Pig uses pig-latin language.	Hive uses HiveQL language.
3.	It was developed by Yahoo.	It was developed by Facebook.
4.	It is used by Researchers and Programmers.	It is mainly used by Data Analysts.
5.	It handles structured and semi-structured data.	It handles Structure data.

[6] Describe Big Data Systems with respect to the 4 characteristics of DW.

Ans. 1. Subject Oriented: A data warehouse is subject oriented because it actually provides information on the specific subject (like a product, customers, suppliers, sales, revenue, etc) not on organization ongoing operation. It does not focus on ongoing operation, it mainly focuses on the analysis or displaying data which help on decision making. Big Data is also subject-oriented, the main difference is a source of data, as big data can accept and process data from all the sources including social media, sensor or machine specific data.

2. Time-Variant: The data collected in a data warehouse is actually identified by a particular time period. As it mainly holds historical data for an analytical report.

Big Data has a lot of approaches to identified already loaded data, a time period is one of the approaches on it. Big data mainly processing flat files, so archive with date and time will be the best approach to identify loaded data. But it has the option to work with streaming data, so it not always holding historical data.

3. Non- Volatile: In DW previous data never erase when new data added to it. This is one of the major features of a data warehouse. As it totally different from an operational database, so any changes on an operational database will not directly impact to a data warehouse.

For Big data, again previous data never erase when new data added to it. It stored as a file which represents a table. But here sometimes in case of streaming directly use Hive or Spark as an operation environment.

4. Distributed file system: Processing of huge data in Data Warehousing is really time-consuming and sometimes it took an entire day to complete the process.

This is one of the big utilities of Big Data. HDFS (Hadoop Distributed File System) mainly defined to load huge data in distributed systems by using map reduce program.

WORKS CITED:

Break out discussions

Lecture discussions

https://cloudxlab.com/assessment/displayslide/345/nosql-cap-theorem

https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html