# IST722: Class Exercise 10

**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: | Bhavya Shah |
| Your Email: | bhshah@syr.edu |

## Instructions (Refer Unit 10)

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 leads to better decision making for businesses. This is because more data gives businesses a more complete picture of their operations and environment. With more data, businesses can see trends that they might not have noticed with smaller datasets, and they can make more confident predictions about the future. For example, a retailer with more data can make better decisions about which products to stock next season.

Data warehouses are designed to store large amounts of data, which makes them ideal for data-driven decision making for businesses.

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

Ans - The CAP theorem states that distributed systems can only guarantee two out of three properties: consistency, availability, and partition tolerance. For example, a distributed database can be consistent and available, but it may not be partition tolerant. This means that if a node in the database goes down, the database may not be able to provide access to all the data.

The CAP theorem is applicable to any distributed system that stores data. It is important for system designers to understand the CAP theorem when designing and implementing distributed systems, so that they can make informed decisions about the trade-offs between consistency, availability, and partition tolerance.

[3] Examine "Schema on Read?" as it relates to Relational Databases and Big Data.

Ans - Schema on read is a data modeling approach that allows data to be stored in a database without a predefined schema. This is in contrast to schema on write, which requires that the schema be defined before the data is stored.

Schema on read is often used with big data, because it can be difficult to know the schema of the data in advance. For example, if you are collecting data from social media, schema on read allows you to store the data without having to define the schema, and then define the schema later when you have a better understanding of the data.

Schema on read can also be used with relational databases. However, it is not as common, because relational databases are designed to work with predefined schemas. Schema on read can be useful with relational databases when the schema is not known in advance, or when the schema needs to be changed frequently.

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

Ans – 3Vs of Big Data are:
1. Volume– Sheer size of data is too large to be processed by a single system.
2. Velocity– The rate at which new data arrives is to frequent to be processed by a single system.
3. Variety– Data are unstructured or semi-structured, so compute resources must add structure before it can be used.

The maximum number of Vs I found is 42Vs of Big Data and Data Science.

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

Ans – 3 differences between Pig and Hive are:
1. Pig is a high-level scripting platform for data processing, while Hive is a data warehousing and SQL-like querying tool for structured data in Hadoop.
2. Pig uses a procedural language called Pig Latin, while Hive uses a declarative language similar to SQL for querying data.
3. Pig is more flexible for custom data processing tasks, while Hive is optimized for querying large datasets using a familiar SQL syntax.

**WORKS CITED:**

- Shafer, T. (2020, November 11). The 42 V's of Big Data and Data Science | Elder Research. Elder Research. **https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/**
- Class slides and Professor Fudge slides.