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: | Yunhan Zhang |
| Your Email: | yzhan297@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.

**I think humans play a very important role in the data-drive decision making process. Humans are needed for planning, accessing and efficient decision making as they have the domain role required for the entire process. Humans have a thing called emotional intelligence for making decisions which machines lack. If user decision needs to be taken, a human is able to identify with the user and understand the real situation rather than machine does.**

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

**The CAP theorem states that a distributed system can only provide two of three properties simultaneously: consistency, availability, and partition tolerance.**

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

**The schema-on-read concept is the counterbalance to the schema-on-write construct. The database schema is created when the data is read. The data structures are not applied or initiated before the data is ingested into the database; it is created during the ETL process.**

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

**The 3Vs are volume, variety and velocity.**

**The 5Vs are velocity, volume, value, variety and veracity.**

**The 10Vs are Volume, Variety, Velocity, Veracity, Validity, Value, Variability, Venue, Vocabulary, and Vagueness.**

**The max number I got is 42Vs.**

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

**1. Hive Component is used mainly by data analysts whereas Pig Component is generally used by Researchers and Programmers.**

**2. Hive Component is used for completely structured Data whereas Pig Component is used for semi structured data.**

**3. Hive Component has a declarative SQL language whereas Pig Component has a procedural data flow language.**

WORKS CITED:

**Professor Fudge’s video and slide.**