Lisa Coombs
February 11, 2025
Ex 5.1 Intro to Big Data

Task 1: What's the difference between structured and unstructured data? Can you give examples that you've encountered for both types?
Structured data is data that is formatted into tables, rows, and columns and is ready for analysis. Unstructured data has no consistent format and usually comes in the form of text messages, emails, photos or videos.
Examples:

Structure Data	**Unstructured Data**
Working with property auction data and downloaded 2024 auctioned property into excel. All the information was neatly put into columns with headers and rows. I can clean and analyze the data immediately.	Needed to find written proof for submission of payments made during a 6 month period. Scrolling was like trying to find a needle in a haystack, even with key words.

Task 2: Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is? What characteristic of big data relates to the question of trustworthiness?
Big Data is inherently untrustworthy because it is essentially created by humans. Humans can be biased, deceptive and dishonest. These human characteristics combined with machine errors can lead to false conclusions when this data is analyzed. Transparency in collecting data is paramount. **Veracity** is the characteristic of trustworthiness in big data.

Task 3: Assume that you receive a table containing customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you've learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.

1. Look at the Data:	If there is a data table, review it to understand the structure, data types and the specific issues present (missing or incomplete values and inconsistent formatting)
2. Remove Duplicates:	Identify and remove duplicate records
3. Handle Missing Values:	Locate missing values and decide on a strategy: **Imputation:** fill in missing value(s) with mean, median or mode **Removal:** remove the columns and/or rows with excessive missing values if it won't negatively impact the overall analysis
4. Standardize Formatting:	Reformat for consistency (i.e. dates, currency) Convert datatypes as necessary Correct inconsistencies (i.e. CA and California)
5. Normalize Data:	Standardizing the numerical values for comparability
6. Validate Data:	Verify through validation checks
7. Document Changes:	Keep a detailed record of all changes made for transparency and reproducibility

Lisa Coombs
February 11, 2025
Ex 5.1 Intro to Big Data

Task 4: Can you describe tools such as Hadoop and Apache Spark and their role in big data? What do they do and how do they work?

	HADOOP	APACHE SPARK
Role in Big Data	open-source framework designed for storing and processing large datasets across distributed computing environments.	a unified analytics engine designed for large-scale data processing.
What it does	**Storage:** uses the Hadoop Distributed File System (HDFS) to store data across multiple nodes in a cluster. **Processing:** uses programming model called MapReduce to process large datasets in parallel.	**In-Memory Computing:** processes data in memory **Versatility:** can support batch processing, real-time streaming, machine learning, and graph processing.
How it works	**HDFS:** Data is divided into blocks and distributed across different nodes. Each block is replicated to ensure fault tolerance. **MapReduce:** breaks down task into smaller sub-tasks (Map) to process in parallel. Combines results (Reduce) for final output. **Components:** NameNode (manages metadata and directory structure) and DataNode (stores actual data).	**In-Memory Processing:** Data is loaded into memory & processed, reducing need for disk I/O operations. **Components:** Spark Core (basic functionality), Spark SQL (structured data processing), Spark Streaming (real-time data processing), MLlib (machine learning), and GraphX (graph processing). **Integration:** Spark can run on Hadoop clusters and access data stored in HDFS, making it compatible with existing Hadoop infrastructure.

Task 5: How has the application of analytics to big data led to new discoveries and innovation? Can you give some examples?

INDUSTRY	NEW DISCOVERIES AND OR INNOVATION
Refugee Support	**Migration Patterns:** use mobile phone and social media posts to track refugee migration patterns. Humanitarian organizations are able to provide timely assistance and plan for future needs. **Sentiment Analysis:** Analyzing social media data helps understand the perceptions and sentiments of host communities towards refugees and can guide communication strategies to foster better integration.
Food Security	**Agricultural Analytics:** monitor crop health, predict yields, and optimize farming practices. Ensures food security in vulnerable regions by improving agricultural productivity and reducing waste. **Supply Chain Optimization:** Analyzing data from the entire food supply chain helps identify inefficiencies and bottlenecks, ensuring that food reaches those in need