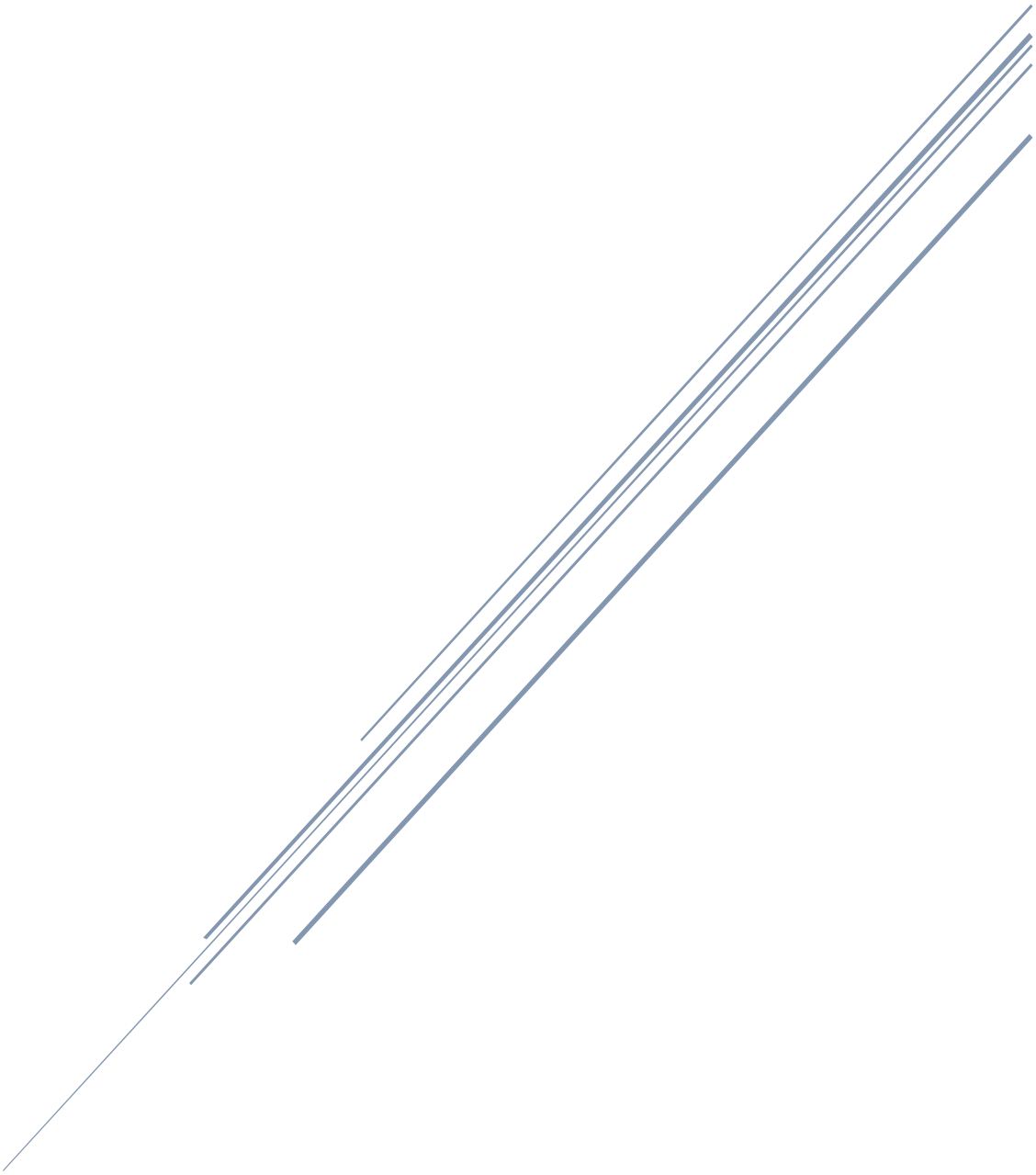
ITRW 321

SU 5 Homework



25948083
Leonard Wassenaar

Big Data Analytics and NoSQL:

Big Data:

3 V's

The volume, velocity and variety characteristic's that is displayed by sets of data.

- Volume-Quantity of data that must be stored.
- Velocity-Speed in which data are going into the system.
- Variety-Difference in the structure of data.

Additional V's

- Variability-The difference in the meaning of data that occurs over a timespan.
- Veracity-The correctness of the data.
- Value-Ask questions like "is the data useful?"
- Visualization-Data must be presented to aid decision making.

Hadoop:

Used for the physical storage of Big Data. The main components of Hadoop include Hadoop distributed File System and MapReduce.

- MapReduce-Distributes data over the processing of data over distributed data.
- Hadoop Distributed File System(HDFS)-The coordination of technology for reliable data distribution on a big cluster of commodity servers.

NoSQL:

Non-relational database approach in data management. The categories of most NoSQL databases are key-value database, document databases, column-oriented, and graph databases.

Key-value database:

Data is stored in key-value pairs, the key value must be known by the distributed database management system. The value data can be of any kind. The key-value database is very fast when data is not dependent.

Document database:

Store the data in key-value pairs. The data on the value side is encoded in a document. This is done by tags, like XML and JSON.

Column oriented database:

Sort data in key-value pairs, where the value side is composed of a set of columns, which are key value pairs.

Graph databases:

Present data as nodes, connected with edges with a certain property. The node is the same as an instance of an entity and the edges represents the association among the nodes. Nodes and edges have properties known as attributes.

NewSQL:

This type of database includes the features of a relational database management system and NoSQL databases.

Data analytics:

Data analytics is a subset of business intelligence functions which provides advanced data analysis tools that extract knowledge form business data. Data analytics can be separated in predictive and explanatory analytics.

- Predictive analytics- Creates models that predict future outcomes by using data that already exists.
- Explanatory analytics- Are geared towards the discovering and explanation of data characteristics and their relationships.

Data mining vs Predictive analytics:

Data mining	Predictive analytics
Automates the analysis of operational data to discover previous data characteristics that is unknown, relationships, dependencies and trends. Four phases: • Data preparation • Data analysis • Classification • Acquisition of knowledge • Prognosis	Use the information generated in the data-mining phase to construct advanced models with high accuracy.