SE ZG522: Big Data Systems

07 BD Analytics And Systems

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

March 19, 2022 ONLINE (WILP @ BITS-Pilani Jan-May 2022)

http://ktiwari.in/bds

BD Analytics: Requirements, Approaches, Techniques

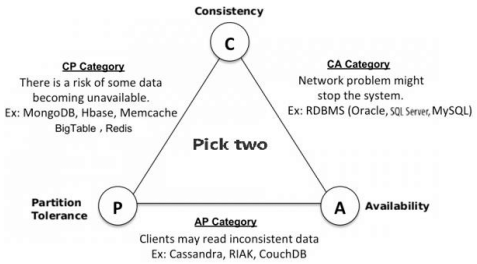
- Date is structured + unstructured (5Vs)
- One could be interested in Analysis or Analytics
- Analytics type: Descriptive, Prescriptive
- Tools: Hadoop, Hive, Storm, Cassendra, Spark, MangoDB

Hadoop - HDFS, Map/Reduce. Spark - Resilient Distributed Dataset (RDD) in memory, Kafka - Pub/Subscriber source/sink connector streaming. Cassandra - Distributed db. mando-db - noSql - not only - key value

BIG Data Systems: Characteristics

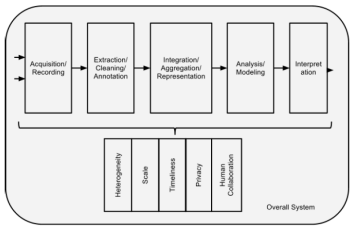
- Failure Tolerance:
  continue operation without interruption when components fail
- Reliability:
  correct output and performance, attack prevention
- Availibility:
  uptime, five 9s (99.999%)
- Consistency:
  agreement (or same) value in multiple copies. [Strict] [Eventual]

CAP Theorem

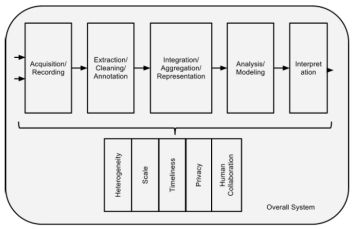


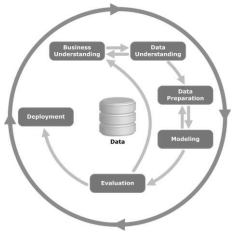
Consistancy, Availability and Partition Tolerance

Big Data Lifecycle



Big Data Lifecycle





Data Acquisition, Data Extraction, Validation and Cleaning, Data Loading, Data Transformation, Data Analysis and Visualization.

Thank You!

Thank you very much for your attention!

Queries ?