Big Data Analytics in SE

1. Why is it important
   1. Extract useful information from large data sets
      1. Quickly
      2. Effectively
   2. Business decisions
      1. Strategic
      2. Tactical
      3. Operational
2. Why is it different
   1. Relational has tight restriction
      1. Poor performance with large data sets
      2. Horizontal scalability issues
      3. Structure defined ahead of time
   2. Big data processing techniques are designed to lift restrictions
      1. Doesn’t guarantee
         1. Atomicity of operations
         2. Consistency of data
      2. Dynamic data structure
3. 3 V’s of Big Data
   1. Volume
      1. Amount of data
   2. Velocity
      1. Speed to be analyzed
   3. Variety
      1. Types of data
4. Structuring big data (NoSQL)
   1. Four main structures
      1. Key-value map
      2. Column-Family
      3. Document
      4. Graph
5. Data and Computing Distribution
   1. Hadoop is commonly used, various components
      1. Distributed file system (HDFS)
      2. Job scheduling and cluster management (YARN)
      3. Distributed computing (MapReduce)
      4. Other Hadoop based projects
         1. Cassandra
         2. Spark
6. Hadoop Distributed File System (HDFS)
   1. Focused on high throughput
   2. Files typically GB to TB in size
   3. Allows for millions of files
   4. Takes a write one read many approach
   5. Easier to move computing to the data over data to the computing
7. Map Reduce
   1. Programming model for processing large amounts of data in parallel on a cluster of machines. Consists of 2 parts
      1. Map
         1. Filters and sorts dataset
      2. Reduce
         1. Performs a summary operation
         2. Processes data and reduce into useable result
8. Advertising
   1. Before big data, companies had to create broad advertisements
   2. With big data adversities can make advertisements tailored to the customer
   3. Analyzing customer demographics to discover market trends
9. Health Care
   1. Personal healthcare device
   2. National health organizations store large amounts of sensitive patient data