June 10, 2018

Course: CIS570 – Business Intelligence
Name: Robert Palumbo
Assignment: Reading Discussions – Week3-Session1
Due Date: Sunday, June 10 @ 11:59pm

1. What is Hadoop? How does it work?

Hadoop is a *newer* data store technology that is based on the concept of distributed computing employing *map-reduce* processing, and was designed as an alternative to traditional relational databases. It is used to both store and further process not only structured data but more importantly massive amounts of *non-structured* data (e.g. social media, video, audio). Hadoop is a member of the *NoSQL (not only SQL)* class of database technologies. It was formally released as *open-source* technology by the *Apache Software Foundation* in 2011.

As mentioned, Hadoop is a *distributed* database system that utilizes lower-cost or *commodity based* computers to form its distributed, networked computing environment. Compared with a traditional RDBMS, the cost of a Hadoop system can be reduced given its open-source format and not having the expense of purchasing expensive proprietary hardware or payment associated licensing fees etc. Thus, it is relatively inexpensive to standup a Hadoop cluster even for personal use.

It consists of two layers. The first is the *Hadoop Distributed File System (HDFS)* layer which are the nodes used to store data. The second layer is the *MapReduce* layer which manages the distributed processing of the data.

Within the HDFS, a node called the *Name Node* manages distributing the data across the other *data* nodes in the cluster. The data (e.g. a blog) is split into chunks (nominally 3 of size 64MB) with each chunk replicated across 3 different nodes. It is the job of the Name Node to keep track of where all the data resides within the system. The replication provides some level of redundancy within the system in case of a node failure.

MapReduce is then used to when it comes time to use the data to solve or find an answer to a specific problem or question. This technique utilizes parallel processing in which the problem to solve is divided up (division of labor) and distributed to the worker nodes. Each worker node has a *Job Tracker* task that manages the process on that node. The worker node carries out its task using the subset of data that it has access to on that node returning its local result when completed. The results obtained from each worker node are then *reduced* resulting in a final answer to the original problem.

https://www.webopedia.com/TERM/H/hadoop_mapreduce.html
https://hadoop.apache.org/

Business Intelligence, Analytics, and Data Science: A Managerial Perspective
Fourth Edition, R. Sharda, D. Delen, E. Turban

1. Compare and contrast end users, business analysts, BI analysts, and data scientists.

Within a business that employs the use of BI as an integral component of its daily operations, there exist a number of various roles that are part of the associated eco-system.

In the context of BI, an *end-user* would be considered anyone that interacts with the system and uses it to obtain an answer to a specific question or problem. For my work at the Judicial Center, end-users routinely contact BI personnel to obtain specific agency related reports such as Judicial Performance Reviews. The end-user has very little (if any) specific knowledge related to the system and its internal workings.

The role or objective of a business analyst in terms of BI is determining and gathering technical requirements for how to best plan for the implement a program or system. The BA with the stakeholders, business partners, and other technologists to ensure all aspects related to project are planned and monitored accordingly, including testing, acceptance, and deployment. Fundamentally the oversee the business-related details for the project.

The role of the BI Analyst becomes more specific and related to the use of the system in terms of being actively involved finding solutions to problems and questions which provide value to those who are the decision makers whether internal or external to the business. Typically, a BI Analyst will utilize the services of the system and data-scientist to gather the necessary intelligence in order to determine a solution providing the most value.

The role of the Data Scientist is primarily to make *sense* out of the massive amount of data that an organization would use for making informed business decisions. This role is very statistics or math centric and typically requires a domain expertise in the business at hand to be able to properly analyze the data for decision making. A DS must be able to dig deep into the data to find hidden trends and insights that even some of the best analytical tools may not uncover. A well-rounded DS will be proficient in both analytical and computational skills being able to find the appropriate meaning within the data and providing an appropriate presentation paradigm to allow value to be extracted from it.

https://www.villanovau.com/resources/business-analysis/business-analyst-job-description/#.WxAUP-6Ut6t
https://expert360.com/blog/role-responsibilities-business-intelligence-analyst/#3
https://www.infoq.com/articles/role-of-a-data-scientist-in-2016