**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 1**

**Part 1**

1. “Thirteen ways to explain Big Data to five years old”

This article is collections of quotes from thirteen industry experts explaining Big Data in simple terms. All of them explains the three “V”s of the Big Data: Volume, Velocity, Variety. Large volume of data that only machine can process, large velocity of data from all gadgets and apps. Different variety of data text, pictures, video, audio…etc.

Some quotes also explain the applications of Big Data about finding different patterns in the data – machine learning. The use cases are ranging from Descriptive Analytics to Predictive and Prescriptive Analytics.

One quote also mentioned importance of the security in the Big Data world. Because organizations collect so much data on everybody and know everything about everybody, it is important we do not allow bad people to access this data.

2. “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights”.

Data Scientist typically spends 50 to 80% of time on Data Wrangling. This process is necessary before can apply modeling algorithms. There are many startup companies offering software to automatically gathering, cleaning, organizing, blending the data from different data sources.

3. “Data Cleansing: Problems and Current Approaches”

This article is on problems and approaches for Data cleaning. Problems can come from single data source or multiple data sources. Data cleaning removes duplicate or incorrect data. Data cleaning approaches uses a process of data analysis, define transformation workflow and mapping rules, verification, transformation, backflow of cleaned data.

4. “Six Core Data Wrangling Activities”

Trifacta defines six core Data Wrangling activities: Discovering, structuring, cleaning, enriching, validating, publishing. It is interesting see more advanced features in a commercial software product would do for Data Wrangling. Before actual data cleaning they use discovering and structuring processes to analyze the dataset.

5. “Beware the Big Errors of Big Data”

The Wired magazine article argues Big Data means any researcher can find spurious correlations between large number of feature variables. The article uses Antifragility to show Big Data Researcher is anti-fragile.

6. “Successful Metrics”

Critical Success Factor (CSF) and Key Performance Indicator (KPI) are examples of important metrics that are used in business world. S.M.A.R.T Metrics includes Specific, Measurable, Achievable, Relevant, Timely. Metric elements contains objective, description, data source, metric owner, reporting period, reporting frequency, present value, previous value, baseline, trend, thresholds.

**Part 2**

Our competitors think using Data would be the next growth engine in health care. From population health, telemedicine, and personal medicine, many new innovations can come from analyze huge volume of Data we currently have. These innovations are using Predictive Analytics and Prescriptive Analytics to achieve better outcome for patients, doctors and providers. We already have large volume of data 150 million of personal records in Health Facts. Now the fun parts are to clean and transform these data into real insights by feeding it into many machine learning models to answer business question and provide value for each of the health care roles (patients, providers, organizations or communities…etc.).

**Part 3**

Big Data is what makes companies like Google and Netflix successful. Type something in Google search box and it will return what is relevant. Google runs Text Mining algorithms on a huge number of documents to categorize relevant documents. Another example is Netflix builds recommendation engine (see Netflix competition) to show user the movies he or she most likely to watch based on many similar users’ past rating.

Big Data processing is difficult because:

* Volume of data. This means bigger clusters, bigger memory requirement.
* Velocity of data. This means constant stream of data; stream processing is difficult.
* Variety of data (structured, unstructured…etc.). How to handle multiple data sources and different types of data are difficult.
* Veracity of data. This means Data Cleaning would be required. The system needs to merge duplicate data and fix missing and incorrect data.

**Part 4**

**Why is it called “Big Data a service”? Justify your answer.**

Web technologies allow companies like Amazon, Google, Facebook, Twitter, Microsoft deploy their Big Data analytics on the web. Software as a service (Saas) deployed in the Cloud enables “Big Data a service”. This web based analytics platform can receive data in large volume, in different format (structured and unstructured), and from many data sources. For examples, Google Analytics and Twitter sentiment analysis. Interactions on Social Media can be traced and business intelligence model trained from interactions between persons. Ecommerce platform can use Social Media data to run text mining, sentiment analysis, recommendation system to better target customers or deliver needed product and service to the customer at right time. This is also true in Healthcare. Genomic data, Probe data, health management data can all be used to create hypothesis testing, knowledge discovery, or innovations.

**Discuss on guidelines for Big Data processing.**

* Guideline 1 Choice of good architecture – Big Data processing is performed on either Batching mode or Streaming model (read time process events).
* Guideline 2 Availability of Analytics

Make data useful requires available of Analytics (machine learning models, visualization on the Big Data platform.

* Guideline 3 Variance in Analytics

Analytics tailored for an application for specific domain with small subset of data.

* Guideline 4 Bring Analytics to Data

Moving large data set to person is not practical, instead migration of analytics is more likely if it is needed.

* Guideline 5 In-memory computation

In-memory analytics that keep data in memory instead of disk is faster.

* Guideline 6 Distributed data storage and in-memory analytics

Distributed data storage is accepted solution for large volume data from different sources. Cloud technology allows saving data and processing data in virtual super computer.

* Guideline 7 Synchrony amount data units

Each data units are associated with requiring analytics. Clusters of data units need to work in synchrony guaranteeing low latency of response.

**Make a list of Big Data tools and note their functionality.**

Open Source tools:

* Hadoop system is consisted of HDFS and MapReduce. Hadoop HDFS stores data on distributed system. Hadoop MapReduce programming model distributes data analytics to many nodes in a cluster to be executed in parallel.
* Hive is a data warehouse system for Hadoop.
* HBase is a NoSQL database system.
* Pig is a high-level data flow language to write map reduce program on Hadoop HDFS.
* Mahout is a Hadoop machine learning library.
* Oozie is a Hadoop job scheduling tool.
* Zookeeper is a Hadoop centralized coordination service.

Commercial tools:

* Microsoft Dryad is a distributed computation execution model that can execute jobs at different locations.
* TIBCO Jaspersoft BI Suite is an open source Eclipsed based data visualization reporting tool.
* Hitachi Pentaho business analytics is another data visualization tool and support many different data stores and support web interface.
* Skytree Server provides advanced analytics using machine learning. It can handle unstructured data, and real time streaming data.
* Tableau data visualization tool we used in Data Visualization class. Tableau server can deploy dashboards to server
* Hortonworks Karmasphere Studio and analyst provides advanced analytics on Hadoop platform.
* Talend Open Studio supports drag and drop interface to develop programs for Hadoop platform for big data processing.
* IBM InfoSphere is a Big Data platform with Hadoop system, also support data warehousing and big data analytics services.