Engineering Big Data - The Next Frontier for Innovation, Competition and Productivity

Snehali Patel[a]

[a] *Toronto, Canada*

Keywords:

1. Introduction

Financial Sector, which always presents the data in Exabytes or Zettabytes a ton of data–and it's the job of Snehali , the Toronto based Data Scientist and Data Integration specialist , to figure out How to handle it , curate it , mine it , wrangle it and present it in a way that the downstream can see the core of the data easily.

Making useful things out of data sometimes requires what seems like an unexpected creative leap–the ability to see how a mandate for research on one track can turn into a product on another. In other words,

The amount of data available today amazes me, Snehali quotes. And her Computer Science background made her realize that this data can be used to make our lives simpler and more efficient. This very fact got her interested in the field of data science and since then she haven't stopped exploring its numerous domains.

She is passionate about data science from business analytics to artificial intelligence, and she thrives on combining Data science with business line to bring productivity to the table.

With an eye for detailed analysis and out of the box thinking to enable stakeholders discover trends buried in their data she is Proficient in using and managing, Cloud subscriptions for actionable insights, to drive business decisions.

2. Aim

Transforming large, unruly data sets into competitive advantages.

3. Material and methods

Data wrangling (sometimes referred to as **data munging**) is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the

data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

The data transformations are typically applied to distinct entities (e.g. fields, rows, columns, data values etc.) within a data set, and could include such actions as extractions, parsing, joining, standardizing, augmenting, cleansing, consolidating and filtering to create desired wrangling outputs that can be leveraged downstream. The recipients could be individuals, such as data architects or data scientists who will investigate the data further, business users who will consume the data directly in reports, or systems that will further process the data and write it into targets such as data warehouses, data lakes or downstream applications.

The terms data wrangling and data wrangler had sporadic use in the 1990s and early 2000s. One of the earliest business mentions of data wrangling was in an article in Byte Magazine in 1997 (Volume 22 issue 4) referencing "Perl's data wrangling services". In 2001 it was reported that CNN hired[“ a dozen data wranglers" to help track down information for news stories.

Data curation is a broad term used to indicate processes and activities related to the organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time and the data remains available for reuse and preservation. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data". In science, data curation may indicate the process of extraction of important information from scientific texts, such as research articles by experts, to be converted into an electronic format, such as an entry of a biological database.

In the modern era of big data the curation of data has become more prominent special when Data is real time, high volume and complex in broad terms, curation means a range of activities and processes done to create, manage, maintain, and validate a component

Data curation is typically user initiated and maintains metadata rather than the database itself. According to the University of Illinois' Graduate School of Library and Information Science, "Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time." The data curation workflow is distinct from data quality management, data protection, lifecycle management and data movement.

The exact curation process undertaken within any organization depends on the volume of data, how much noise the data contains and what the expected future use of the data means to its dissemination. Many high-ranking institutions employ data curators and mass information management systems. Data curators view information storage methodically and from various viewpoints, and as institutions grow their information stacks, they will require more data curation specialists.

January 22, 2018

Data integration involves combining data residing in different sources and providing users with a unified view of them. This process becomes significant in a variety of situations, which include both commercial (such as when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains. Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved.

Consider a web application where a user can query a variety of information about cities (such as crime statistics, weather, hotels, demographics, etc.). Traditionally, the information must be stored in a single database with a single schema. But any single enterprise would find information of this breadth somewhat difficult and expensive to collect. Even [redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

4. [redacted]

[redacted]

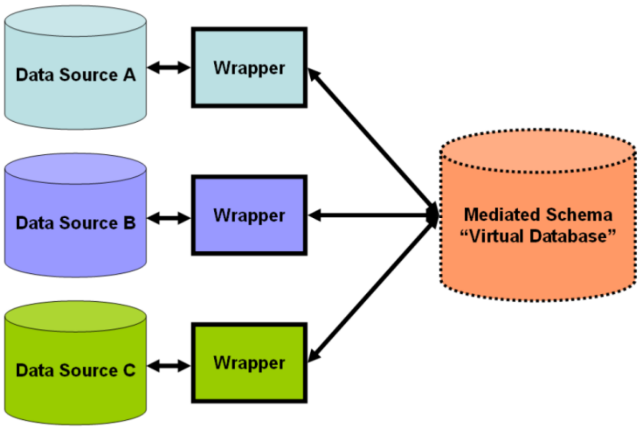

Figure 1:

[redacted].

[redacted].

[redacted].

5. [redacted]

[redacted].– *Kasey Zachari*

[redacted].

6. [redacted]

[redacted].

Author biography

Snehali Patel Data Integration Specialist

January 22, 2018