Semantic Extract-Transform-Load framework for Big Data Integration

Srividya K Bansal
Arizona State University, Mesa, AZ, USA
srividya.bansal@asu.edu

Sebastian Kagemann
Indiana University – Bloomington, IN, USA
sakagema@umail.iu.edu

***Abstract*— Big Data researchers are dealing with the Variety of data that includes various formats such as structured, numeric, unstructured text data, email, video, and audio. The proposed Semantic Extract-Transform-Load (ETL) framework that uses semantic technologies to integrate and publish data from multiple sources as open linked data provides an extensible solution for effective data integration, facilitating the creation of smart urban apps for smarter living. A case study that integrates datasets, using the proposed framework, from various Massive Open Online Courses and Household travel data along with Fuel Economy data is presented.**

Keywords—data integration; linked data; ontology engineering; semantic technologies

Big Data comprises of data consisting of billions to trillions of records of millions of people - all from different sources (e.g. Web, customer contact center, social media, mobile data, sales, etc.). The data is typically loosely structured and is often incomplete and inaccessible. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. Massive amounts of data are available to be harvested for competitive business advantage, government policies, and new insights into a broad array of applications (including healthcare, biomedicine, energy, smart cities, genomics, transportation, etc.). Yet, most of this data is inaccessible to users, as we need technology and tools to find, transform, analyze, and visualize data in order to make it consumable for decision-making [1]. The research community also agrees that it is important to engineer Big Data meaningfully [2]. Meaningful data integration in a schema-less, and complex Big Data world of databases is a big open challenge. Big Data research is usually discussed in the areas of 3V's – Volume (storage of massive amount of data streaming in from social media, sensors, and machine-to-machine data being collected), Velocity (reacting quickly enough to deal with data in near-real time), and Variety (data is in various formats such as structured, numeric, unstructured text data, email, video, audio, stock ticker, etc.). Big Data challenges are not only in storing and managing this variety of data but also extracting and analyzing consistent information from it. Researchers are working on creating a common conceptual model for the integrated data [3].

The method of publishing and linking structured data on the web is called Linked Data. This data is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and it can be linked to from other data sets as well. The Linked Open Data (LOD) community effort has led to a huge data space, with 31 billion Resource Description Framework (RDF) [4] triples, and a W3C specification for data interchange on the web [5]. LOD can be used in a number of interesting Web and mobile applications. Linking Open Government Data (LOGD) project [6] investigates translating government-related data using Semantic web technologies. LOD has gained significant adoption and momentum, though the quality of the interconnecting relationships remains questionable [7]. IBM Smarter City initiative aims at creating cities that are vital and safe for its citizens and businesses. Their focus is on building the infrastructure for fundamental services—such as roadways, mass transit and utilities that make a city desirable and livable. IEEE Smart Cities Initiative brings together technology, government and society to enable smart economy, mobility, environment, living, and governance. Both these initiatives have to integrate and use information from various data sources in addition to setting up the required infrastructure. Government agencies are also increasingly making their data accessible through initiatives such as data.gov to promote transparency and economic growth [8]. We need ways to organize variety of data such that concepts with similar meaning are related through links, while the concepts that are distinct are clearly represented as well with semantic metadata. This will allow effective and creative use of query engines and analytic tools for Big Data, which is absolutely essential to create smart and sustainable environments. Figure 1 shows the future vision of a web portal with Linked Open Urban data integrated and published from various sources and domains. The need to integrate Big Data has been heightened in recent years due to a growing demand and interest in mobile applications for improving quality of life in urban cities. Here is an example where various data sources can be used: a traffic jam that emerges due to an unplanned protest may be captured through a Twitter stream, but missed when examining weather conditions, event databases, reported roadwork, etc. Additionally, weather sensors in the city tend to miss localized events such as flooding. These views of the city combined however, can provide a richer and more complete view of the state of the city, by merging traditional data sources with messy and unreliable social media streams thereby contributing to smart living, environment, economy, mobility, and governance. Such applications rely on Big Data available to the public via the cloud.

As outlined in the latest McKinsey Global Institute report, we're now seeing the global economy beginning to operate in real time [9]. The total value generation for the impact of new data technologies will be measured in trillions of dollars globally according to this report. The National

Academies press published Visionary Grand Challenges of Manufacturing for 2020 that included as one of the challenges – *ability to instantaneously transform information gathered from a vast array of sources into useful knowledge for making effective decisions*. Work has been done on data abstraction and visualization tools for Big Data [10]; analysis of Big Data using various algorithms such as influence-based module mining (IBMM) algorithm, online association rule mining, graph analytics, and provenance analysis support framework all of which are applicable after the data integration phase. Use of semantic technologies has known to improve user interaction with the system, simplified data integration and extensibility, improved search, and improved discovery of knowledge [11].

Figure 1: Linked Open Urban Data portal with integrated data from various sources and domains

Traditionally Data Integration has been defined as the problem of combining data residing at different sources, and providing the user with a unified view of these datasets [13-14]. A data integration system exposes to its users a schema for posing queries. This schema is typically referred to as a mediated schema (or global schema). To answer queries using the various information sources the system needs mappings that describe the semantic relationships between the mediated schema and the schemas of the sources. Two basic approaches have been proposed for this purpose. The first approach, called global-as-view, requires that the global schema be expressed in terms of the data sources. The second approach, called local-as-view, requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema [14-15]. These existing approaches were applicable to relational data models and an integrated global schema was determined manually. The proposed framework in this paper is for integration of linked data that is available on the web.

MOTIVATING SCENARIO:

Consider the following scenario where a typical driver, John gets into his car in the morning and turns on the ignition. A built-in innovative application in the car greets him and asks him if he is going to work based on the time of the day. John responds by saying "yes" and the app replies that the vehicle performance has been optimized for the trip. The built-in system uses the GIS system, road grade data, and speed limits data to create an optimal velocity profile. As John starts driving and is approaching *Recker road* to turn left, the app informs John about a road repair on *Recker road* for a 1-mile stretch on his route up to 3pm that day. The app suggests that John should continue driving and take the next left on *Power road.* John follows the suggestion and answers a text message that he receives from his collaborator. As he is answering the text message, his car drifts into the neighboring lane. The app immediately notifies John of the drift who quickly adjusts his driving. As John approaches his workplace he drives towards *Lot 1* where he usually parks. The app informs John that there are only 2 parking spots open in *Lot 1*. As John is already running late for a meeting, he decides to directly drive to the next parking lot, *Lot 2,* to avoid spending the time looking for the 2 empty spots in *Lot 1*. As John enters *Lot 2* and is driving towards one of the empty spots, he gets too close to one of the parked cars. The app immediately warns John of a collision. John quickly adjusts his car away from the parked cars and parks in an empty spot. The app logs tracking data about John's style of driving on the server for future use.

In order to build apps for automobiles, access to a number of data sets from various sources is required. Some of this is real-time data that is continuously being updated. Data related to traffic, road repairs, emergencies, accidents, driving habits, maps, parking, fuel economy data, household data, diagnosis data, etc. would be required. Various automotive apps could be built that focus on reducing energy consumption, reducing emissions, provide fuel economy guidance that is based on actual vehicle data, road conditions, traffic, and most importantly personal driving habits and ways to improve them. It is important to effectively integrate data such that the data is tied to a meaningful and rich data model that can be queried by these innovative applications.

ETL PROCESS:

The Extract-Transform-Load (ETL) process in computing has been in use for integration of data from multiple sources or applications, possibly from different domains. It refers to a process in data warehousing that extracts data from outside sources, transforms it to fit operational needs, which can include quality checks, and loads it into the end target database, more specifically, operational data store, data mart, or data warehouse. The three phases of this process are described as follows:

- Extract: this is the first phase of the process that involves data extraction from appropriate data sources. Data is

usually available in flat file formats such as csv, xls, and txt or is available through a RESTful client.

- Transform: this phase involves the cleansing of data to comply with the target schema. Some of the typical transformation activities involve normalizing data, removing duplicates, checking for integrity constraint violations, filtering data based on some regular expressions, sorting and grouping data, applying built-in functions where necessary, etc. [12].
- Load: this phase involves the propagation of the data into a data mart or a data warehouse that serves Big Data.

One of the popular approaches to data integration has been ETL as shown in, which describe the taxonomy of activities in ETL and a framework using a workflow approach to design ETL activities. A declarative database programming language called LDL was used to define the semantics of ETL activities. Similarly, there have been other approaches such as UML and data mapping diagrams for representing ETL activities, quality metrics driven design for ETL, and scheduling of ETL activities. A number of tools facilitate the ETL process, namely IBM Infosphere, Oracle Warehouse Builder, Microsoft SQL Server Integration Services, and Informatica Powercenter for Enterprise Data Integration. Talend Open Studio, Pentaho Kettle, CloverETL are open source ETL products. The focus in these existing approaches has been on the design of ETL workflow and not about generating meaningful/semantic data that is important in integration of data from variety of sources. Semantic approaches to ETL technologies have been proposed that use semantic technologies to further enhance definitions of the ETL activities involved in the process rather than the data itself. A tool that allowed semi-automatic definition of inter-attribute semantic mappings, by identifying parts of data source schemas, which are related to the data warehouse schema has been proposed by the research community. This supported the extraction phase of ETL. The use of semantics here was to facilitate the extraction process and workflow generation with semantic mappings.

SEMANTIC ETL FRAMEWORK:

The proposed semantic ETL framework generates a semantic model of the dataset(s) under integration, and then generates semantic linked data in compliance with the data model. This generated semantic data is made available on the web as linked data available for querying, analytics, or used in data-driven innovative apps. The use of semantic technologies is introduced in the Transform phase of an ETL process to create a semantic data model and generate semantic linked data (RDF triples) to be stored in a data mart or warehouse. The transform phase will still continue to perform other activities such as normalizing and cleansing of data. Extract and Load phases of the ETL process would remain the same. Figure 2 shows the overview of activities in semantic ETL. Transform phase will involve a manual process of analyzing the datasets, the schema and their purpose. Based on the findings, the schema will have to be mapped to an existing domain-specific ontology or ontology will have to be created from scratch. If the data sources belong to disparate domains, multiple ontologies will be required and alignment rules will have to be specified for any common or related data fields.

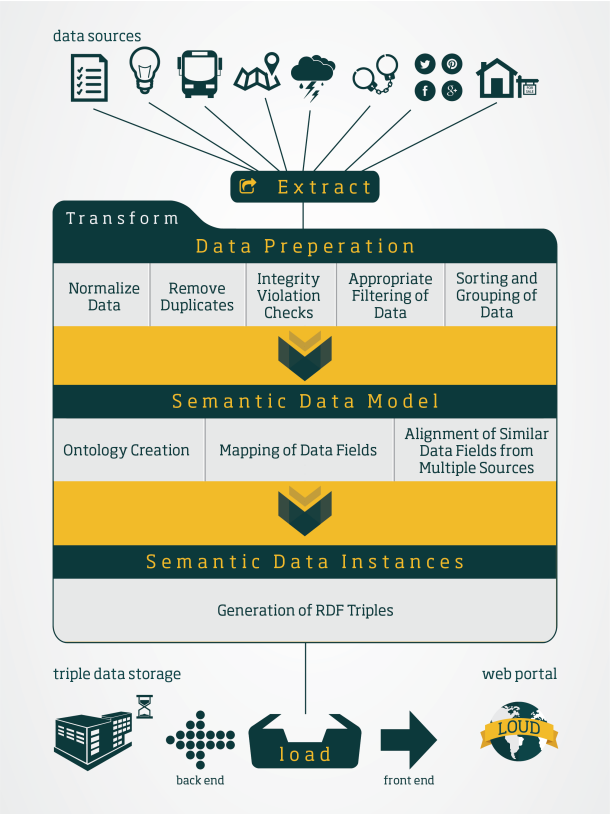

Figure 2: Semantic ETL Framework

TECHNOLOGY STACK:

Semantic web technologies facilitate: the organization of knowledge into conceptual spaces, based on their meanings; extraction of new knowledge via querying; and maintenance of knowledge by checking for inconsistencies. These technologies can therefore support the construction of an advanced knowledge management system. The following Semantic technologies and tools are used as part of our Semantic ETL framework:

- *Uniform Resource Identifier (URI)*, a string of characters used to identify a name or a web resource. Such identification enables interaction with representations of the web resource over a network (typically the Web) using specific protocols.
- *Resource Description Framework (RDF)*, a general method for data interchange on the Web, which allows the sharing and mixing of structured and semi-structured data across various applications. As the name suggests, RDF is a language for describing web resources. It is used for representing information, especially metadata, about web resources. RDF is designed to be machine-readable so that it can be used

in software applications for intelligent processing of information.

- *Web Ontology Language (OWL)* is a markup language that is used for publishing and sharing ontologies. OWL is built upon RDF and an ontology created in OWL is actually a RDF graph. Individuals with common characteristics can be grouped together to form a class. OWL provides different types of class descriptions that can be used to describe an OWL class. OWL also provides two types of properties: object properties and data properties. Object properties are used to link individuals to other individuals while data properties are used to link individuals to data values. OWL enables users to define concepts in a way that allows them to be mixed and matched with other concepts for various uses and applications.
- SPARQL – a RDF Query Language, which is designed to query, retrieve, and manipulate data stored in RDF format. SPARQL allows for a query to consist of triple patterns, conjunctions, disjunctions, and optional patterns. SPARQL allows users to write queries against data that can loosely be called "key-value" data, as it follows the RDF specification of the W3C. The entire database is thus a set of "subject-predicate-object" triples.
- Protégé Ontology Editor - Protégé is an open-source ontology editor and framework for building intelligent systems. It allows users to create ontologies in W3C's Web Ontology Language. It is used in our framework to provide semantic metadata to the schema of datasets from various sources.

CASE STUDY:

A case study was conducted to assess the effectiveness of the proposed Semantic ETL framework using public datasets on Educational data from Massive Open Online courses (MOOC), National Household travel survey data and EPA's Fuel Economy data.

MOOC Course Data:

Data from 3 MOOC providers Coursera, edX, and Udacity was integrated using the proposed framework for this study. Coursera is the largest MOOC provider in the world with 7.1 million users in 641 courses from 108 institutions as of April 2014. These courses span 25 categories including 4 subcategories of computer science. Course data is retrievable via Coursera's RESTful course catalog API and returned in JSON. edX is a premier MOOC provider with more than 2.5 million users and over 200 courses as of June 2014. edX courses are distributed among 29 categories, many of which overlap with Coursera's. edX does not provide an API for accessing their course catalog, however, edX's entire platform is currently open-source. Udacity is a vocational course-centric MOOC provider with 1.6 million users in 12 full courses and 26 free courseware as of April 2014. The majority of Udacity courses are within the field of computer science. Udacity does not provide an API for accessing their course catalog data.

Household Travel Data:

National Household Travel survey (NHTS) data, published by the U.S. Department of Transportation, has been used in this study. This data was collected to assist transportation planners and policy makers who needed transportation patterns in the United States. The dataset consists of daily travel data of trips taken in a 24-hour period with information on the purpose of the trip (work, grocery shopping, school dropoff, etc.), means of transportation used (bus, car, walk, etc.), how long the trip took, day of week when it took place, and additional information in case a private vehicle was used. This dataset has been used by research community to study relationships between demographics and travel, correlation of the modes of transportation, amount, or purpose of travel with the time of the day and day of the week. This dataset was integrated with Fuel Economy data to discover the performance of private vehicles used by members of the household.

Fuel Economy Data:

The U.S. Environmental Protection Agency's (EPA) data on Fuel Economy of vehicles provides detailed information about vehicles produced by all manufacturers (make) and models. It provided detailed vehicle description, mileage data for both city drive and highway drive, mileage for different fuel types, emissions information, and fuel prices. EPA obtained this data as a result of vehicle testing conducted by EPA's National Vehicle and Fuel Emissions Lab and the manufacturers.

Semantic Data Model generation

A common semantic data model was designed and developed to integrate Household Travel and Fuel Economy data using Protégé for ontology engineering. This model comprised of primary classes Person, Household, Vehicle at the top most level. All the fields in the datasets were modeled as classes, subclasses, or properties and connected to the primary classes via suitable relationships. The ontology had both object properties and data properties depending on the type of data field being represented. Figure 3 shows the high-level semantic data model with the primary classes and properties and their relationships. The semantic data model for integrating MOOCs data was created by extending Schema.org that has generic creative work described and organized as a hierarchy of types, each associated with a set of properties. Learning Resource Metadata Initiative (LRMI) specification adds properties to the *CreativeWork* object of Schema.org including the time it takes to work through a learning resource (timeRequired), the typical age range of the content's intended audience (typicalAgeRange), etc. This existing vocabulary for *CreativeWork* was used as a base type extended to add *Course*, *Session*, *Category* and their associated properties drawn from MOOC datasets. Figure 4 shows the high-level semantic data model. The complete ontology can be obtained from the project website.

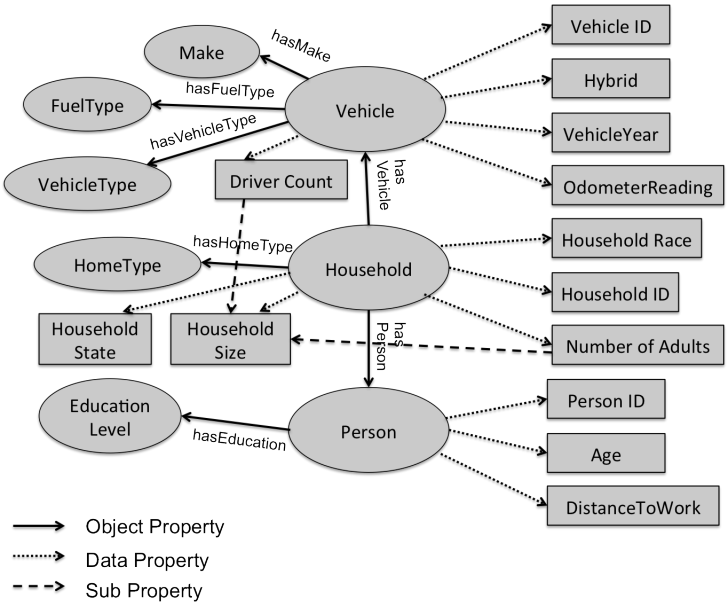

Figure 3: Semantic data model for household travel and fuel economy data

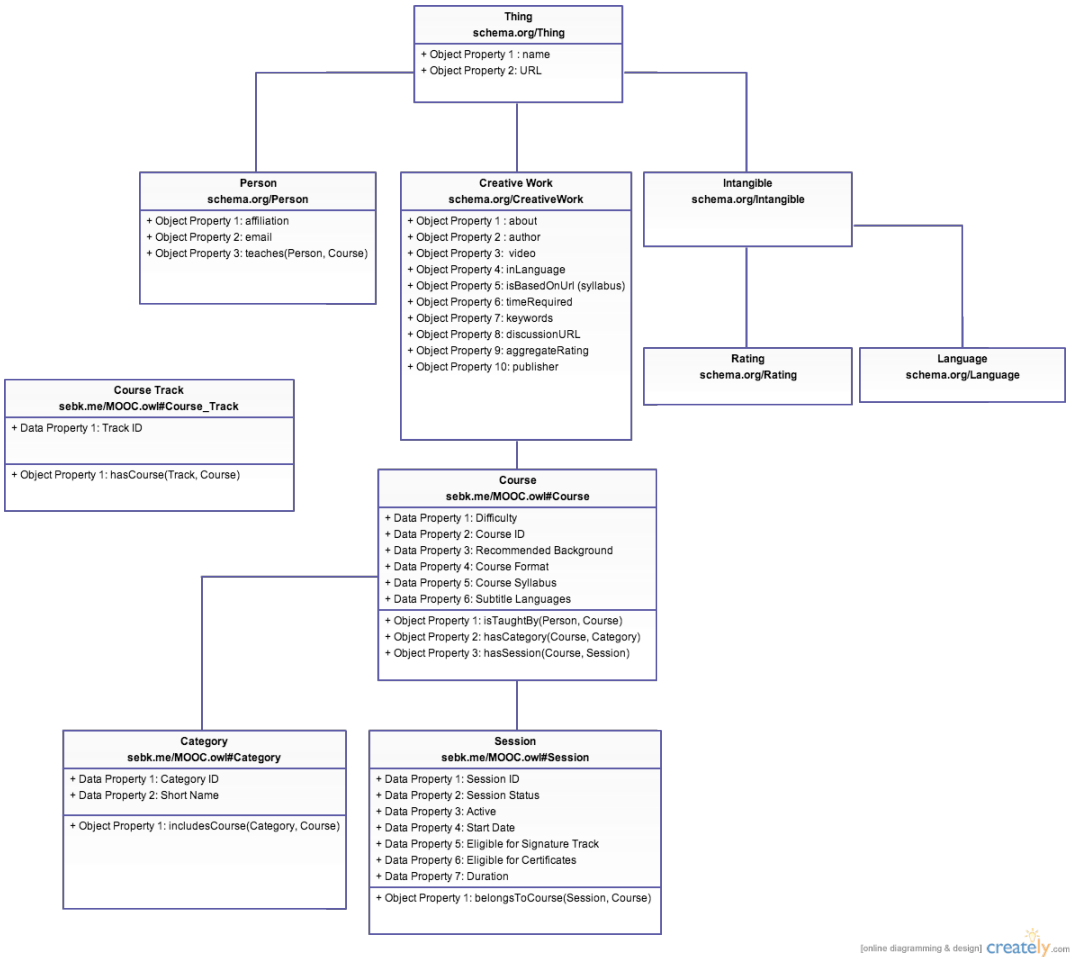

Figure 4: Semantic Data Model for MOOC data set

Semantic Instance Data generation

Household travel and fuel economy datasets were obtained from in *CSV* format that was converted into *RDF* triples using *Oxygen XML* editor. A web crawler was written using *Scrapy* Crawler framework to obtain MOOC datasets in *JSON* format. A total of 30,000 triples were generated that can be accessed from the project website.

Semantic Querying

The integrated datasets were queried using RDF Query language *SPARQL*. The Java open source framework for building Semantic Web and Linked data applications, Apache Jena, was used for executing SPARQL queries on travel and fuel economy data. An application was built that loads the ontologies and data, checks for conformity of data with the vocabulary, and then run SPARQL queries against the data. A sample list of queries is provided in Table 1. Linked data generated from MOOCs was stored into a stand-alone Fuseki server (a sub-project of Jena) that provides an HTTP interface to RDF data. A sample SPARQL query is shown below:

```
PREFIX mooc: <http://sebk.me/MOOC.owl#>
PREFIX schema: http://schema.org/
SELECT * WHERE {
   ?course rdf:type mooc:Course.
   ?course schema:name ?iname.
   FILTER (regex(?iname, "calculus", "i")).
}
```

Table 1: Sample Queries for integrated household travel and fuel economy data

#	Query
1	Most popular car in U.S
2	Most popular car in AZ
3	Number of diesel cars in a particular census division
4	Estimated Annual Cost of regular gasoline per census division
5	Estimated Annual Cost of gasoline per household
6	Gas Guzzler tax per household

A MOOC aggregator web application called MOOCLink was developed that queries this integrated data set to provide consolidated information about courses being offered, upcoming courses, allows users to search for courses, and compare courses. This web application is currently under testing. A screenshot of a course detail page of the web application is shown in Figure 5. Further details of this ongoing project can be updated from the project website.

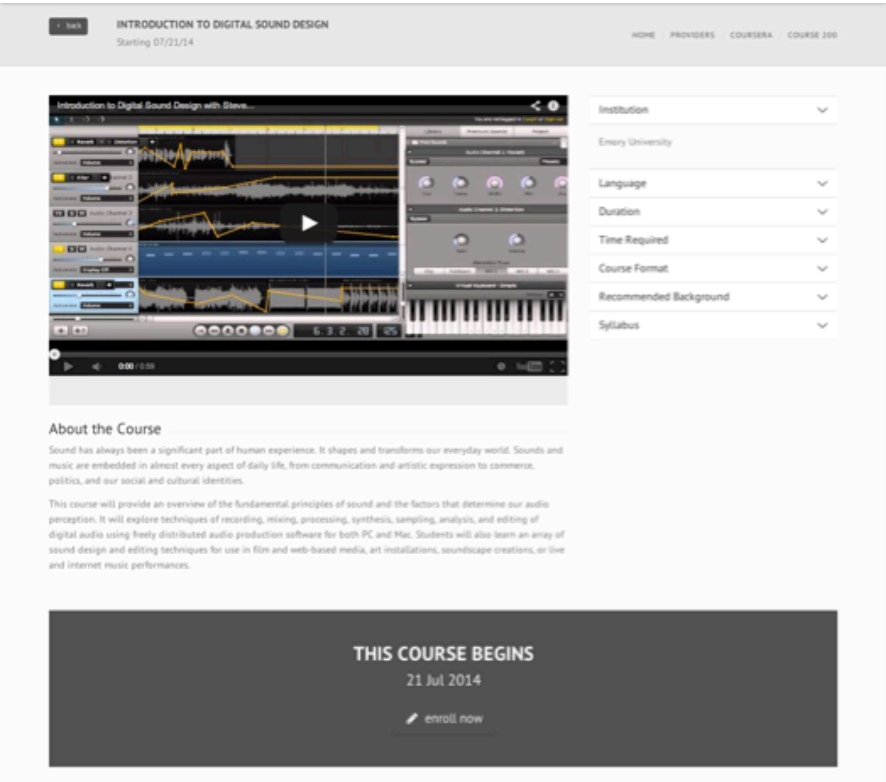

Figure 5: Screenshot of Course detail page of the MOOC aggregator web application

Evaluation

Current tools that facilitate the ETL process focus in these existing approaches has been on the design of ETL workflow and *not* about generating meaningful/semantic data that is important in integration of data from variety of sources. The Semantic ETL framework focuses on providing semantics to various data fields thereby facilitating richer data integration.

LESSONS LEARNED:

Integration of data from various heterogeneous sources into a meaningful data model that allows intelligent querying and creation of innovative applications is an important open issue in the area of Big Data. The case study conducted showed great potential for the proposed Semantic ETL framework in producing Linked Open Data that would facilitate the creation of data-driven innovative apps for smart living. The Semantic Data Model generation process used in the case study comprises of manual analysis of data, ontology engineering, creation of linked data (RDF triples). One of the challenges of this project is ontology engineering that needs a fairly good understanding of the data from different sources. A human expert has to perform this step and it is often a time-consuming process that involves a deep understanding of the data sets under integration. There is a need for algorithms that automatically (or with minimal human intervention) generate semantic data models. Based on this requirement the following are some future research directions:

- Establishing a process to identify existing ontologies for data sets under integration
- Extending an existing ontology with relevant properties, classes, and relationships or creation of a new ontology
- Generate alignment rules between multiple ontologies that might exist for individual data sets
- Generate alignment rules between the Semantic data model that is created and existing well-known ontologies such as FOAF, DBPedia, DublinCore, and Wordnet to produce semantically rich Linked Open Data.

Semantic ETL process can be introduced into existing open source ETL software tools such as CloverETL - an Open Source Engine that is a Java library and does not come with any User Interface components. It provides powerful data transformation and ETL features that are used in the commercial Clover edition. CloverETL is a rapid, end-to-end data integration solution popular for its usability and intuitive controls, along with its lightweight footprint, flexibility, and processing speed.

Future directions for this project will also involve the creation of interesting and innovative applications (web and/or mobile) in various domains such as automotive, aerospace, healthcare, education to list a few.

REFERENCES:

[1] E. Kandogan, M. Roth, C. Kieliszewski, F. Ozcan, B. Schloss, and M.-T. Schmidt, "Data for All: A Systems Approach to Accelerate the Path from Data to Insight," in *2013 IEEE International Congress on Big Data (BigData Congress)*, 2013, pp. 427–428.

[2] C. Bizer, P. Boncz, M. L. Brodie, and O. Erling, "The Meaningful Use of Big Data: Four Perspectives – Four Challenges," *SIGMOD Rec.*, vol. 40, no. 4, pp. 56–60, Jan. 2012.

[3] A. Azzini and P. Ceravolo, "Consistent Process Mining over Big Data Triple Stores," in *2013 IEEE International Congress on Big Data (BigData Congress)*, 2013, pp. 54–61.

[4] P. Hayes and B. McBride, "Resource description framework (RDF)," 2004. [Online]. Available: http://www.w3.org/TR/rdf-mt/. [Accessed: 28-Feb-2014].

[5] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International journal on semantic web and information systems*, vol. 5, no. 3, pp. 1–22, 2009.

[6] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, and others, "TWC LOGD: A portal for linked open government data ecosystems," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 3, pp. 325–333, 2011.

[7] S. Dastgheib, A. Mesbah, and K. Kochut, "mOntage: Building Domain Ontologies from Linked Open Data," in *2013 IEEE Seventh International Conference on Semantic Computing (ICSC)*, 2013, pp. 70–77.

[8] F. Lecue, S. Kotoulas, and P. Mac Aonghusa, "Capturing the Pulse of Cities: Opportunity and Research Challenges for Robust Stream Data Reasoning," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[9] "Big data: The next frontier for innovation, competition, and productivity | McKinsey & Company." [Online]. Available: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. [Accessed: 12-Jul-2014].

[10] S. K. Bista, S. Nepal, and C. Paris, "Data Abstraction and Visualisation in Next Step: Experiences from a Government Services Delivery Trial," in *2013 IEEE International Congress on Big Data (BigData Congress)*, 2013, pp. 263–270.

[11] C. C. Wang, D. A. Hecht, P. C. Sheu, and J. J. Tsai, "Semantic Computing and Drug Discovery," 2013.

[12] P. Vassiliadis, A. Simitsis, and E. Baikousi, "A Taxonomy of ETL Activities," in *Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP*, New York, NY, USA, 2009, pp. 25–32.

[13] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: the teenage years," in *Proceedings of the 32nd international conference on Very large data bases*, 2006, pp. 9–16.

[14] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 233–246.

[15] A. Calì, D. Calvanese, G. De Giacomo, and M. Lenzerini, "Data integration under integrity constraints," in *Seminal Contributions to Information Systems Engineering*, Springer, 2013, pp. 335–352.

Srividya Bansal *is an Assistant Professor in Ira A. Fulton Schools of Engineering at Arizona State University Her primary research focuses on semantics-based approaches for Big Data Integration, Web service description, discovery & composition, and tools for outcome-based instruction design in STEM education. She is also interested in Software Engineering Education research that focuses on experimenting various delivery models in project-centric courses. She received a PhD in Computer Science from the University of Texas at Dallas. She is a member of the IEEE Computer Society. Contact her at srividya.bansal@asu.edu.*

Sebastian Kagemann *is a student in the department of Computer Science at Indiana University – Bloomington where he is pursuing B.S. in Computer Science. His academic interests include mobile development, distributed computing and cognitive science. Contact him at sakagema@umail.iu.edu.*