CS 519- Big Data Exploration and Analysis - Spring 2014

Project Report

Project Title: Patterns of Heterogeneity in Databases

Team Members:

Pratik Krishnamurthy krishnap@onid.oregonstate.edu

Catharina Vijay vijayc@onid.oregonstate.edu

Nitin Jogee Bella Subramanian subraman@onid.oregonstate.edu

Motivation:

If everyone used the same database structure for representing the same information, things would have been “simpler”. However, we often use heterogeneous data sources where the same information is represented in different ways. Having homogeneous structures for information from different data sources become the need for databases. In this project, we try to proceed towards design independence by analyzing the differences in the structures of such heterogeneous datasets in several domains.

Problems:

A lot of data is getting generated every single day. Organizations store data the size of which is rising at an exponential rate. It does not make sense to generate data without benefiting from it. We need to analyze data to find meaning in it. ‘Heterogeneity’ is a major problem concerning big data analytics. In simple terms and as explained above, heterogeneity means different format or representation for the same data.

Goals:

The goals of this project comprise of:

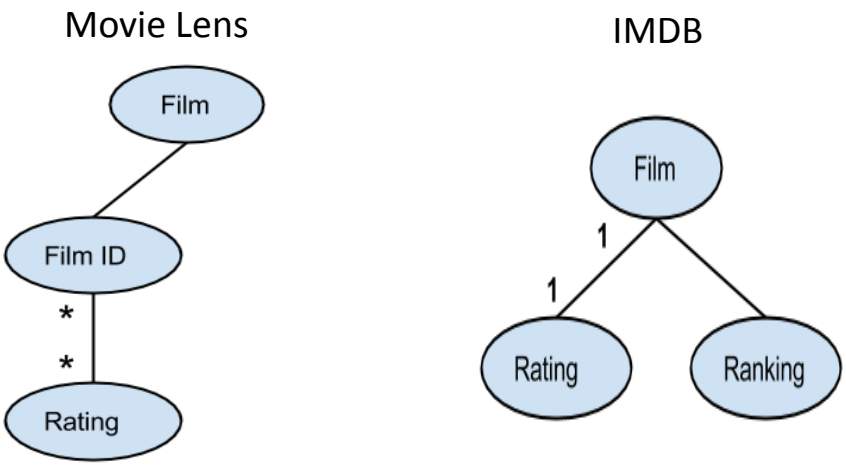
- Analyzing the common differences in representation of same information across different real world datasets including semantic and structural differences.
- Getting closer to design independence. Design independence means that a user, who gets an answer for a query over one schema, should not have to change the query for getting the same answer over another schema.

Observations:

Domain 1: Movies

Datasets:

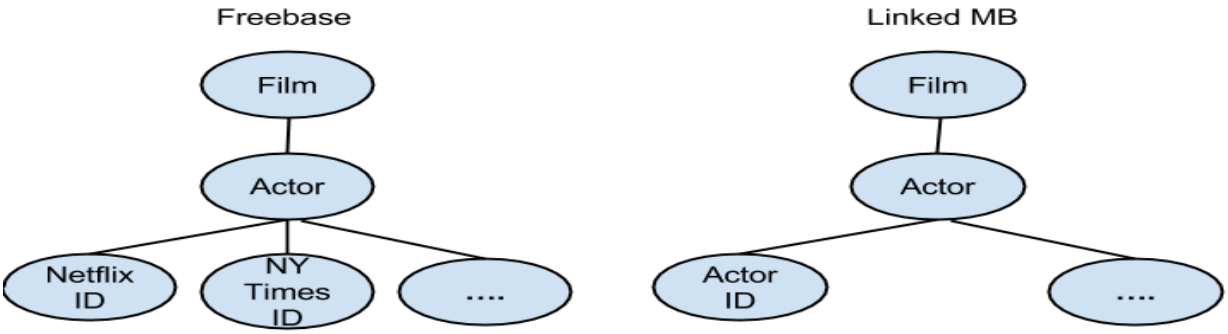
1. MovieLens
2. IMDB
3. Freebase
4. Linked Movie Database = (Freebase + Other datasets)



MovieLens: One Flew Over the Cuckoo's Nest– 1193 (Film ID) – 5 (rating),
James and the Giant Peach– 661 (Film ID) – 3 (rating),
James and the Giant Peach– 661 (Film ID) – 5 (rating)

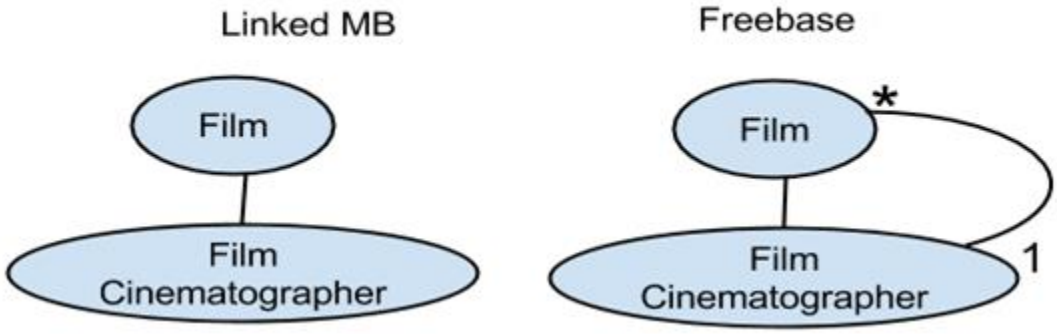
IMDB: 12 Angry Men – 8.1 (rating) – 157 (rank),
One Flew Over the Cuckoo's Nest – 8.1 (rating) – 162 (rank)

In MovieLens dataset, 'Rating' serves as both rating and ranking while in IMDB, Rating and Ranking are separate nodes. Though rating and ranking are different, they could be used interchangeably either intentionally or un-intentionally.
Also, MovieLens has many user ratings for a particular film. The overall rating assigned by MovieLens can be shared by many movies. Hence, we can see an n-to-n relationship between Film and Rating. IMDB has ratings for the top 250 movies where each film has only one rating and the rating is unique. Hence, we can see a 1-to-1 relationship between Film and Rating in IMDB. This structural difference could arise if the perspective is different. Eg. Since the IMDB dataset consists of the top 250 movies, each film would need to have a unique rating. In the case of MovieLens, the data in the dataset is not aligned for such a purpose.



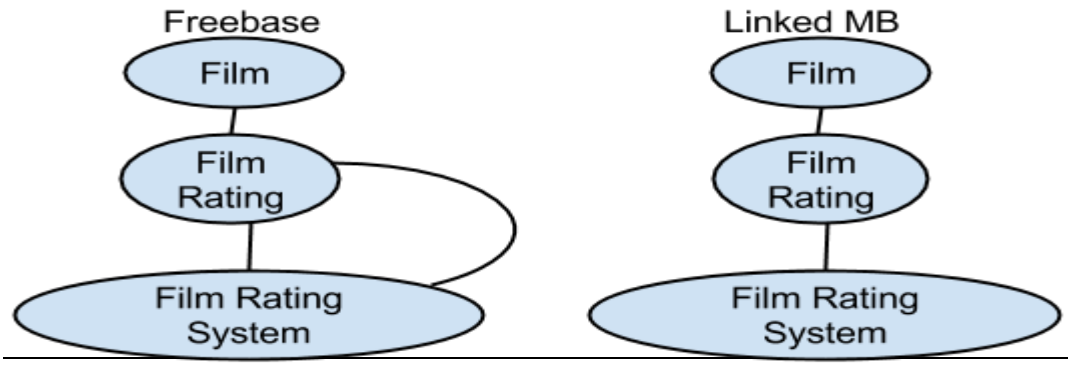
Freebase: Resident Evil Apocalypse – Milla Jovovich – 47184 (Netflix ID),
Linked MB: Resident Evil Apocalypse - Milla Jovovich – 30535 (Actor ID)

Both Freebase and Linked MB contain different nodes under Actor. While the Actor in Freebase contains Netflix ID and NY Times ID, the one in Linked MB contains Actor ID. Freebase seems to be more flexible when it comes to accessing a movie based on different IDs like Netflix ID, NY Times ID, etc.



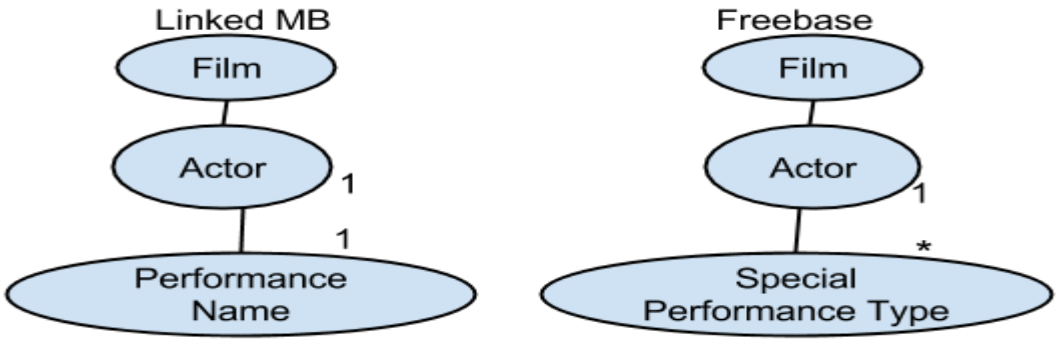
Freebase: Apocalypse Now – Vittorio Storaro – 1900, Apocalypse Now, Ishtar,….
Linked MB – Apocalypse Now – Vittorio Storaro

In Linked MB, Film cinematographer contains the links pointing to the films he/she has worked in. We can visualize link(s) between Film and Film Cinematographer and can call it a 1-to-n relationship. In the case of Freebase, no such relationship exists.



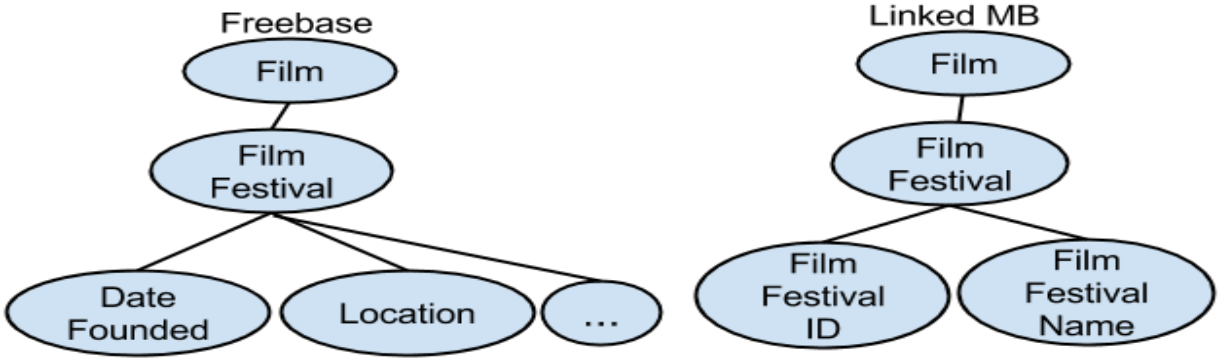
Freebase: UK – PG – British Board of Film Classification – UK: Uc, UK: PG, UK:12,…
Linked MB: Film Rating - PG

In Freebase, Film Rating System contains the link(s) pointing to Film Rating that we have indicated with a curve. This is not the case with Linked MB. We suspect that the integration of Freebase with other datasets resulting in Linked MB, could have caused the change in the structure.



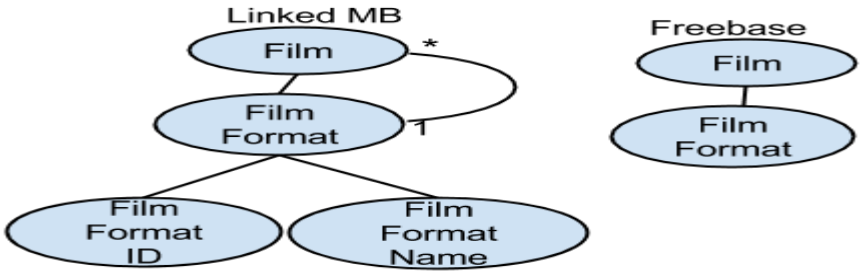
Linked MB: Stand and Deliver – Mark Everett – performance name
Freebase: 2 Fast 2 Furious – Paul Walker – Performance names (more than one will be listed wherever applicable).

Linked MB has a 'Performance Name' node that has a 1-to-1 relationship with Actor. In the case of Freebase, 'Special Performance Type' has an n-to-1 relationship with Actor. The structure of Freebase looks more realistic because an actor could have more than 1 performance names in a film. But one thing to note is that Linked MB is derived from Freebase. Due to integration of Freebase with other datasets, the structure of Linked MB might have been modified resulting in a structure that is seen above.



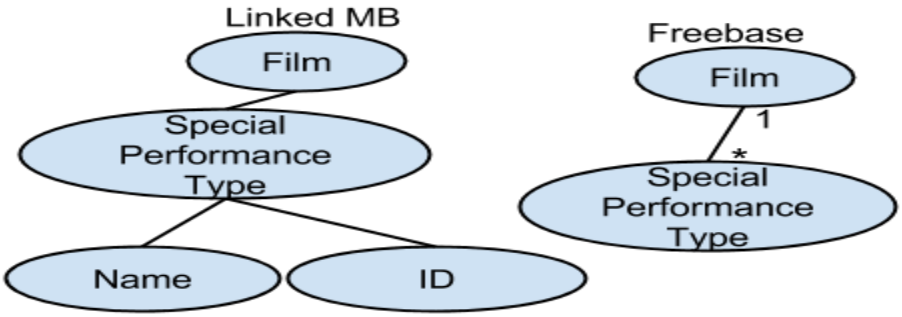
Freebase: 2007 Berlin International Film Festival – Date founded, Berlin
Linked MB: 2005 Berlin Film Festival – 184, 2005 Berlin Film Festival

The nodes under Film Festival in both Freebase and Linked MB are different hinting that the purpose served by Film Festival in both datasets are different. The Film Festival under Freebase has sub-nodes that denote the description for one festival while that in Linked MB database accommodates multiple film festivals.



Linked MB: Moulin Rouge – DVD Video (with format ID 4 and formant name DVD-Video), 35 mm (with format ID 47 and format name 35 mm film).
Freebase: Amelia - DVD

The Film Format node in Freebase has no sub-nodes while that in Linked MB has sub-nodes. Also, based on our observation, Film Format in Linked MB has a 1-to-n relationship with Film while this is not the case in Freebase.



Linked MB: Cameo Appearances – Cameo Appearance, 11
Freebase: Amelia – many special performance types (wherever applicable)

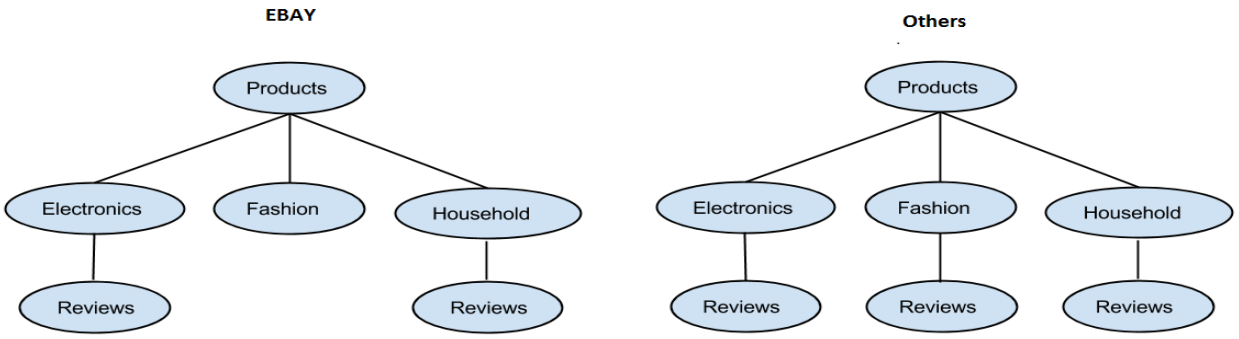
The Special Performance Type node in Freebase has no sub-nodes while that in Linked MB has sub-nodes. Also, based on our observation, Special Performance Type in Freebase has n-to-1 relationship with Film while this is not the case in Linked MB. As explained in an earlier example, Freebase accommodates multiple performance types for Film in the same node while Linked MB contains Name and ID for each Special Performance Type.

Domain 2: Products

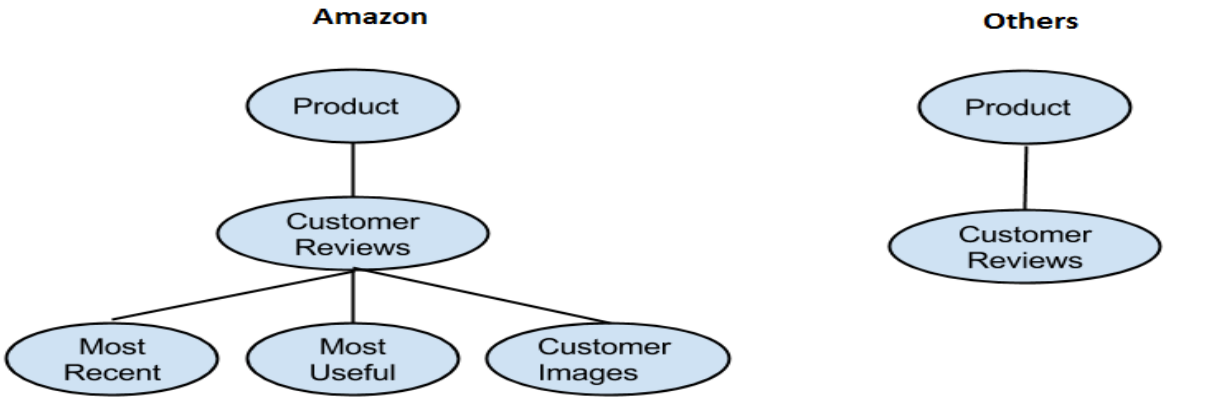
Datasets:

1. Amazon
2. Walmart
3. eBay
4. Sears

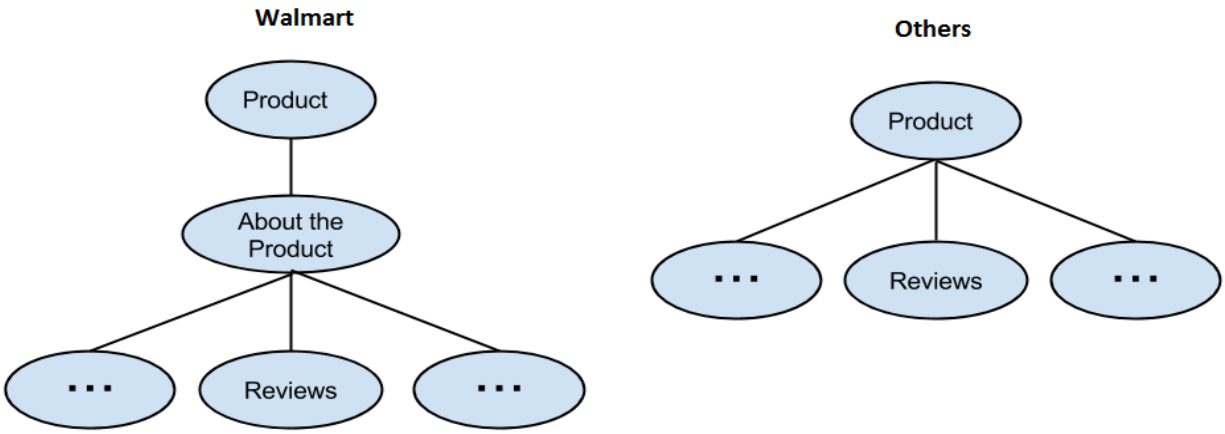
We analyzed three categories under each of the above datasets: Electronics, Fashion and Household.



eBay does not have customer reviews under Fashion while the other 3 datasets have reviews for each product category. The above structure comes as a result of analysis of a subset of the dataset.



Amazon categorizes Customer Review as Most Recent and Most Useful while the other sites don't do that. They usually display the reviews with the most recent review at the top. Compared to other datasets, Amazon tries extracting more meaning from the data by labeling the data as Most Useful, Most Recent, etc.

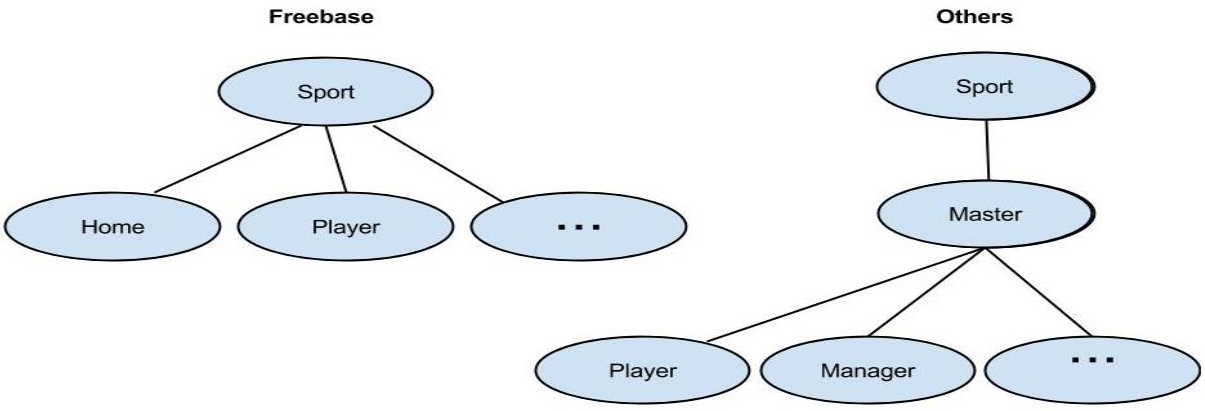


Walmart has customer review under ‘About the product’ while the other sites have the Review node directly under Product. Walmart represents the relationship between Product and Reviews in the form of a node while the other datasets express the relationship as an edge.

Domain 3: Sports

Datasets:

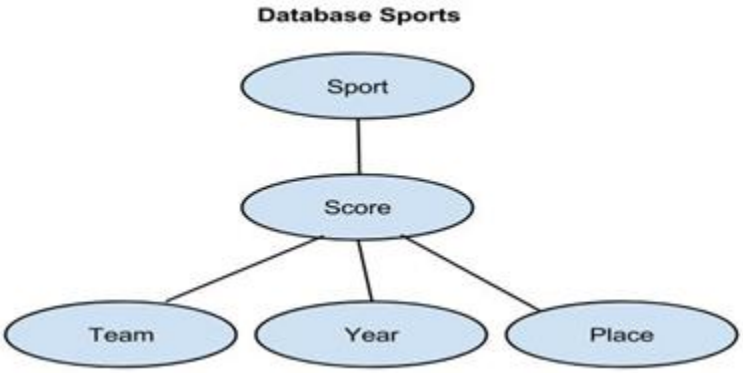
1. ESPN
2. DatabaseSport.com
3. Freebase
4. OpenSourceSports.com



This is a structural differences and hence, we could not come up with examples.

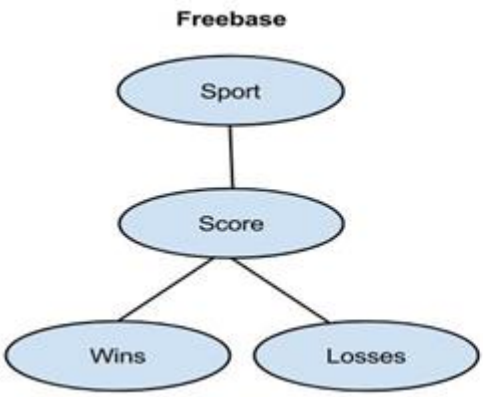
ESPN has a Home node that does not contain any other node. But OpenSource Sports has a Master node which has other sub-nodes like Player, Manager etc. From a developer point-of-view, the Master node could be useful to access other nodes (in

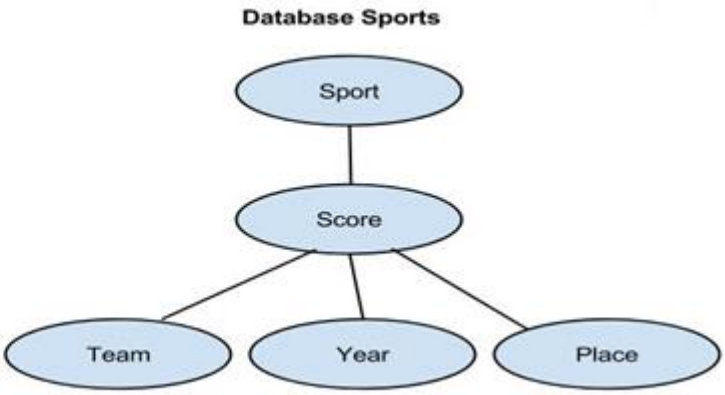
other words, querying) because the complexity of the queries could reduce with each keyed attribute in one single relation.



ESPN: NBA – Various scores
Database Sports: Basketball – Los Angeles, 105 at Utah, 101 Box

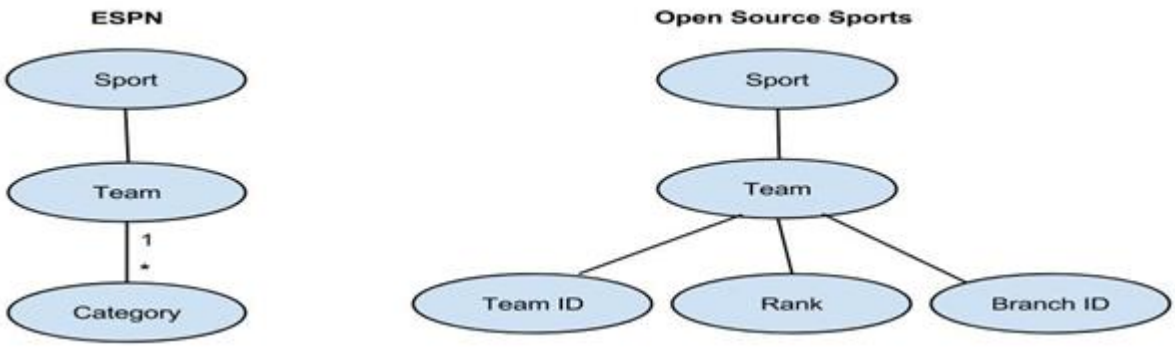
Scores appear unbranched in ESPN database while it is branched in Database Sports.





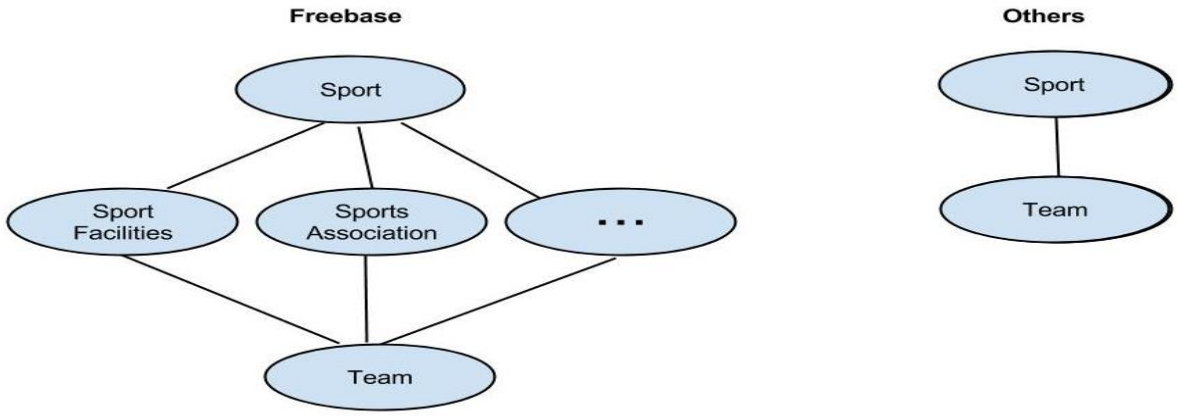
Freebase: NBA – Sports Team Season Records – 9 wins, 0 losses
Database Sports: Basketball – Los Angeles, 2001, Utah

Freebase represents scores in terms of wins and losses while in the case of Database Sports, the score is represented in terms of other parameters. Hence, the way of approaching Scores is different in these datasets.



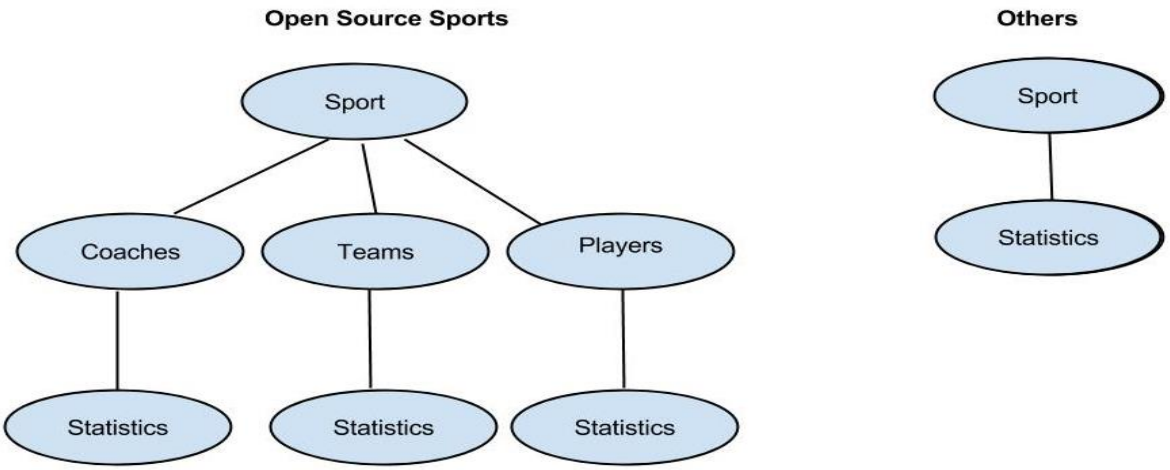
Freebase: NBA – Team – Atlantic, Pacific, Central, South West, …
OpenSource Sports: Basketball – Team – BOS, 5, BOS

In case of ESPN, there is only 1 sub-node under Teams called Category. Also, Teams have 1-to-n relationship with Category. In case of OpenSource Sports, there are many sub-nodes nodes under Team and the nodes are different compared to the node under Team for ESPN, hinting that these databases follow different approaches to represent Team.



Structural difference observed and hence, could not come up with example.

In Freebase, Team is a sub node under many other nodes of Sports but the other datasets have Team only under Sport.



Structural difference observed and hence, could not come up with example.

OpenSource Sports contain statistical attributes for Coaches, Teams and Players while the other datasets have a separate node called Statistics under Sport that contains statistics for all the nodes under Sport.

Conclusion:

Based on the observations stated above, we conclude the following:

- We have observed a lot of differences in the datasets including different ways of representing the same attribute, missing attributes, etc. We have ignored reporting such differences for this project.
- For e-commerce sites where we observed different product categories, we found that the representation of data under each categories is uniform. Eg. In Electronics, if reviews are present under ‘About the product’ for one product, the other products in the same category also follow the same structure.
- We observed structural differences across datasets of a domain where many nodes were related to each other. Eg. In a domain like Films where there are a lot of relations between the cast, crew and other entities, we could spot structural differences across various datasets for that domain.
- In general, people have different ways of representing the same information. To make the data analysis process easier, it is necessary to go for design independent structures.
- There is a lot of analysis of data needed to achieve design independence.

Resources:

1. **Film domain:**
   a. Freebase – http://www.freebase.com/film
      (Looked at both the schema and the graph through the web UI)
   b. Linked Movie Database – http://linkedmdb.org
      (Looked at the RDF through web UI)
   c. IMDB – http://www.imdb.com (Dataset shared with you)
   d. MovieLens – Datasets shared with you

2. **Product domain:** All product links are shared separately with you
   a. Amazon – http://www.amazon.com
   b. eBay – http://www.ebay.com
   c. Sears – http://www.sears.com
   d. Walmart – http://www.walmart.com

3. **Sports domain**
   a. ESPN – http://www.espn.com (Analyzed web structure)
   b. Freebase – http://www.freebase.com/sports (Analyzed web based graph)
   c. Database Sports – http://www.databasesports.com (Analyzed web structure)
   d. OpenSource Sports – http://www.opensourcesports.com (Analyzed web structure)