Big Data Integration

Avigdor Gal

Technion – Israel Institute of Technology

Tutorial Outline

- Big data
- Data integration
- Applications of big data integration
- Current challenges and future research directions

Big data is a game changer

- From Theory to Systems: empirical evaluation counts
- From Systems to Data: large scale empirical evaluation counts

Data Volume: No Longer the Size of a Teacup

Volume				

Table : Big Data Cross Table

Big data may be a single dataset with a lot of data

Data Volume: No Longer the Size of a Teacup

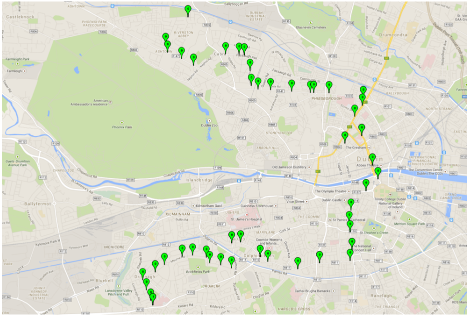


Table : Big Data Cross Table

Big data may be a single dataset with a lot of data

Data Velocity: Replacing a Teacup with a Tea Hose

Volume				
Velocity				

Table : Big Data Cross Table

Big data may be data that rapidly changes

Data Velocity: Replacing a Teacup with a Tea Hose

Table : Big Data Cross Table

Big data may be data that rapidly changes

Data Velocity: Replacing a Teacup with a Tea Hose

Table : Big Data Cross Table

Big data may be data that rapidly changes

Data Velocity: Replacing a Teacup with a Tea Hose

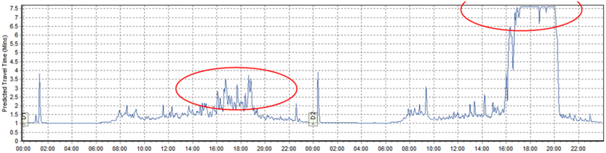


Table : Big Data Cross Table

Big data may be data that rapidly changes

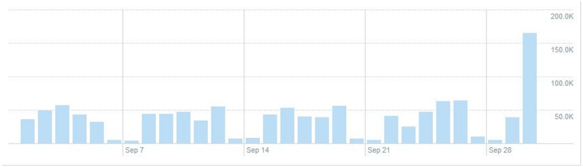
Data Variety: When One Tea Type is Just not Enough

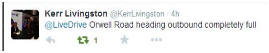
Volume				
Velocity				
Variety				

Table : Big Data Cross Table

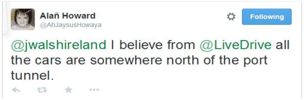
Big data may be a small dataset with many different schemata

Data Variety: When One Tea Type is Just not Enough





Kerr Livingston @KerrLivingston · 4h
@LiveDrive Orwell Road heading outbound completely full



Alaň Howard
@AhJaysusHowaya

@jwalshireland I believe from @LiveDrive all the cars are somewhere north of the port tunnel.

Table : Big Data Cross Table

Big data may be a small dataset with many different schemata

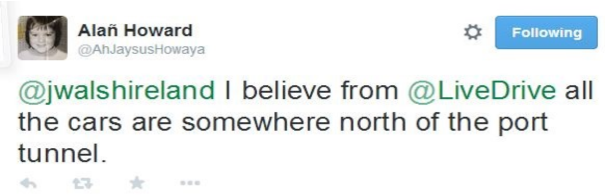
Data Veracity: Is it Coffee or Black Tea with Milk?

Volume				
Velocity				
Variety				
Veracity				

Table : Big Data Cross Table

Big data may be data with varying levels of trustworthiness

Data Veracity: Is it Coffee or Black Tea with Milk?



Alañ Howard
@AhJaysusHowaya

@jwalshireland I believe from @LiveDrive all the cars are somewhere north of the port tunnel.

Table : Big Data Cross Table

Big data may be data with varying levels of trustworthiness

Data Gathering: where and when to expect the fountain to burst

	Gathering			
Volume				
Velocity				
Variety				
Veracity				
	Signal and Event Processing			

Table : Big Data Cross Table

Data Gathering: where and when to expect the fountain to burst

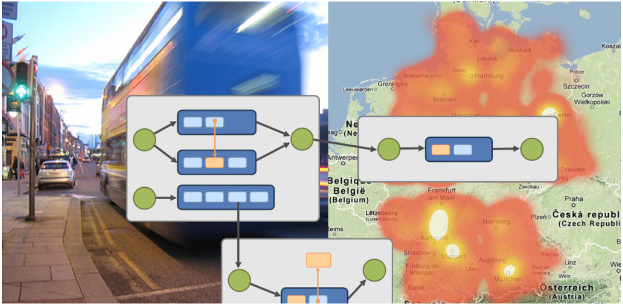


Table : Big Data Cross Table

Data Management: Not your typical DBA anymore

	Gathering	**Managing**		
Volume				
Velocity				
Variety				
Veracity				
		Cloud Computing, NoSQL, NewSQL		

Table : Big Data Cross Table

Data Analytics: When Data Analysis Explodes Multi-Dimensionally

	Gathering	**Managing**	**Analyzing**	
Volume				
Velocity				
Variety				
Veracity				
			Data & Process Mining ML, IR, NLP	

Table : Big Data Cross Table

Data Visualization: The Machine Offering to Mankind

	Gathering	**Managing**	**Analyzing**	**Visualizing**
Volume				
Velocity				
Variety				
Veracity				
				User Experience

Table : Big Data Cross Table

Data Visualization: The Machine Offering to Mankind

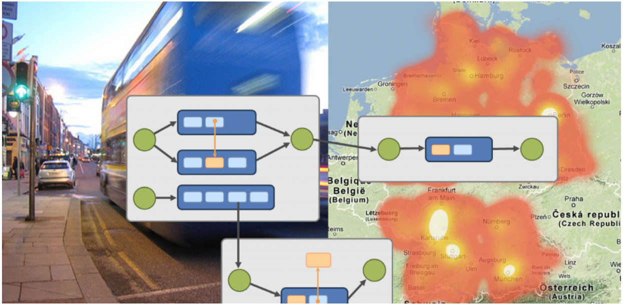


Table : Big Data Cross Table

Big Data Cross Table

	Gathering	Managing	Analyzing	Visualizing
Volume				
Velocity				
Variety				
Veracity				

Table : Big Data Cross Table

What is Data Integration?

- Data Integration is the task of integrating multiple data sources into a single data source.
- Data Integration is a management task in the Big Data Cross Table.
- Two major tasks of data integration are schema matching and entity resolution.

Schema Matching

What is Schema Matching?

- Ancient history: heterogeneity of schemata
    - Different DBAs, different names
    - Granularity matters
- Schema matching is the process of creating attribute correspondences among multiple schemata

Schema Matching

What is Schema Matching?

- Ancient history: heterogeneity of schemata
  - Different DBAs, different names
  - Granularity matters
- Schema matching is the process of creating attribute correspondences among multiple schemata

Existing Work

- Formal Models: uncertain schema matching
- Algorithmic & Heuristic solutions: string, value, structure-based
- Empirical benchmarks: University applications, Web forms, Ontology matching competition (OAEI)

Textbook Example for Schema Matching

Id	name	ZIP	Income
r_1	Green	51519	30K
r_2	Green	51518	32K
r_3	Peter	30528	40K
r_4	Peter	30528	40K

Table : SM Simple Example

Id	firstName	lastName	Address	Salary
r_1	John	Green	CARTER LAKE IA 51519	30,000
r_2	Sarah	Green	CARTER LAKE IA 51518	32,000K
r_3	Peter	Smith	CLEVELAND GA 30528	40,000
r_4	Peter	Smith	CLEVELAND GA 30528	40,000

Table : SM Simple Example 2

Entity Resolution

What is Entity Resolution?

- Real world data is dirty
    - Typographical errors and missing values
    - Different date formats and terminology
    - Multiple representations of the same real-world object
    - Multi-dimensional data aspects: temporal, spatial, ...
- ER is the process of determining when different entity representations refer to the same entity.

Entity Resolution

What is Entity Resolution?

- Real world data is dirty
  - Typographical errors and missing values
  - Different date formats and terminology
  - Multiple representations of the same real-world object
  - Multi-dimensional data aspects: temporal, spatial, ...
- ER is the process of determining when different entity representations refer to the same entity.

Existing work

- Formal Models and Languages
- Algorithmic solutions
- Comparative empirical analysis of solutions: FEBRL

Textbook Example for Entity Resolution

Id	name	ZIP	Income
r_1	Green	51519	30K
r_2	Green	51518	32K
r_3	Peter	30528	40K
r_4	Peter	30528	40K
r_5	Gtee	51519	55K
r_6	Howard	51519	30K

Table : ER Simple Example

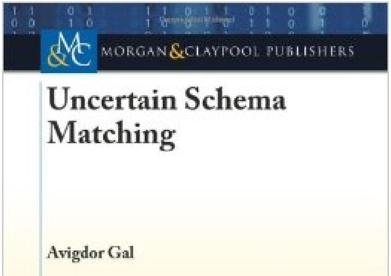
Big Data + Data Integration = Big Data Integration

	Gathering	**Managing**	**Analyzing**	**Visualizing**
Volume		ER		
Velocity		ER		
Variety		SM		
Veracity		SM & ER		

Table : Big Data Cross Table

Big Data + Data Integration = Big Data Integration

	Gathering	Managing	Analyzing	Visualizing
Volume		ER		
Velocity		ER		
Variety		SM		
Veracity		SM & ER		



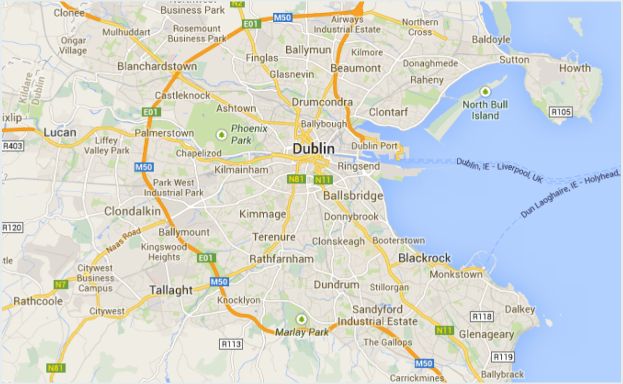
MORGAN & CLAYPOOL PUBLISHERS

Uncertain Schema Matching

Avigdor Gal

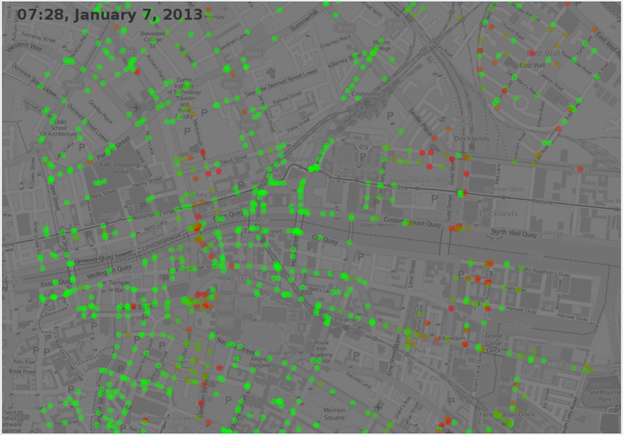
Big Data Integration: not Your Typical Data Integration Anymore

Urban Traffic Management



Big Data Integration: not Your Typical Data Integration Anymore

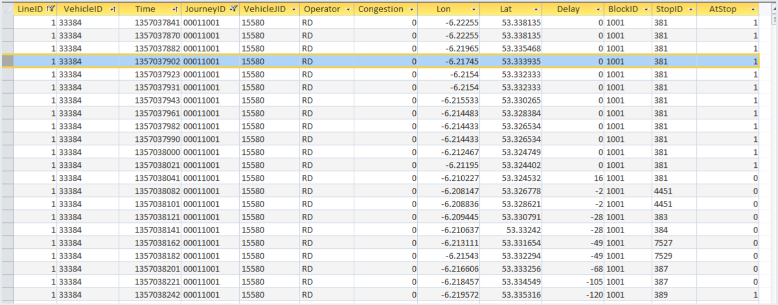
Traffic Flow



07:28, January 7, 2013

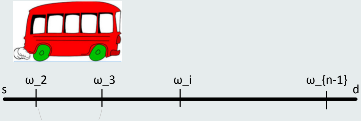
Big Data Integration: not Your Typical Data Integration Anymore

Bus Log



LineID	VehicleID	Time	JourneyID	VehicleJID	Operator	Congestion	Lon	Lat	Delay	BlockID	StopID	AtStop
1	33384	1357037841	00011001	15580	RD	0	-6.22255	53.338135	0	1001	381	1
1	33384	1357037870	00011001	15580	RD	0	-6.22255	53.338135	0	1001	381	1
1	33384	1357037882	00011001	15580	RD	0	-6.21965	53.335468	0	1001	381	1
1	33384	1357037902	00011001	15580	RD	0	-6.21745	53.333935	0	1001	381	1
1	33384	1357037923	00011001	15580	RD	0	-6.2154	53.332333	0	1001	381	1
1	33384	1357037931	00011001	15580	RD	0	-6.2154	53.332333	0	1001	381	1
1	33384	1357037943	00011001	15580	RD	0	-6.215533	53.330265	0	1001	381	1
1	33384	1357037961	00011001	15580	RD	0	-6.214483	53.328384	0	1001	381	1
1	33384	1357037982	00011001	15580	RD	0	-6.214433	53.326534	0	1001	381	1
1	33384	1357037990	00011001	15580	RD	0	-6.214433	53.326534	0	1001	381	1
1	33384	1357038000	00011001	15580	RD	0	-6.212467	53.324749	0	1001	381	1
1	33384	1357038021	00011001	15580	RD	0	-6.21195	53.324402	0	1001	381	1
1	33384	1357038041	00011001	15580	RD	0	-6.210227	53.324532	16	1001	381	0
1	33384	1357038082	00011001	15580	RD	0	-6.208147	53.326778	-2	1001	4451	0
1	33384	1357038101	00011001	15580	RD	0	-6.208836	53.328621	-2	1001	4451	0
1	33384	1357038121	00011001	15580	RD	0	-6.209445	53.330791	-28	1001	383	0
1	33384	1357038141	00011001	15580	RD	0	-6.210637	53.33242	-28	1001	384	0
1	33384	1357038162	00011001	15580	RD	0	-6.213111	53.331654	-49	1001	7527	0
1	33384	1357038182	00011001	15580	RD	0	-6.21543	53.332294	-49	1001	7529	0
1	33384	1357038201	00011001	15580	RD	0	-6.216606	53.333256	-68	1001	387	0
1	33384	1357038221	00011001	15580	RD	0	-6.218457	53.334549	-105	1001	387	0
1	33384	1357038242	00011001	15580	RD	0	-6.219572	53.335316	-120	1001	389	1

Bus Model



Big Data Integration: not Your Typical Data Integration Anymore

Big Data Challenges

Volume: 23 Million records per month ($\sim 4GB$)

Velocity: 770,000 new records per day (an event each 2-6 seconds)

Variety: Homogeneous

Veracity: GPS locations

Challenges of Big Data Integration

Big data

The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades. (Hal Varian, Google's Chief Economist)

Data integration

Data integration has been the basis of data understanding and processing for many years now. With big data joining in, the impact of data integration is not diminishing. Rather, it changes shape while remaining dominant.

Challenges of Big Data Integration

Challenges

Volume — Compute data integration faster, by using parallelization.

Velocity — Create incremental computation methods for data integration.

Variety — Extend evaluation models to support data integration with minimal or no human input in the loop.

Veracity — Quantified uncertainty management for data integration.

Lecture Outline
Big Data
Data Integration
Big Data Integration
Challenges of Big Data Integration

Thank You

Avigdor Gal

Technion – Israel Institute of Technology