UvA Data Service: Towards an Unified Architecture for Scientific Data Aggregation

Introduction

Cross-disciplinary discoveries are brought by interdisciplinary research and exchange of data/experiments. The OSDC platform is a growing venue for such data. However, one obstacle is the lack of a tool to discover data content and to allow data exchange, independent of the store format. To this end, we want to provide two types of tools: one that deals with categorizing the data content, by using generic semantics about it (what field it comes from, how it was obtained: experiment, benchmark, measurement, simulation); and another which deals with non-functional aspects of the data, e.g. type of database, geo-location.



Research Questions

How can we create a tool that allow users throughout the world to easily share their data within research communities?

Can this tool unify the different data management alternatives (databases) so the users need not be concerned about the details of the data management but rather about the data itself?







Pedro D. Bello-Maldonado¹, Ana Oprescu², Paola Grosso², Cees de Laat² ¹Florida International University, ²Universiteit van Amsterdam

Architecture



MySQL	Database

INIYOUL DALADASE												Ivry O &						
	ID Coord X Coord Y Intensity												ID	Coord >				
	S1	10		25		60			Name	X		Υ	Value		S1	10		
	S2	15		21		53		\rightarrow	S1	10		25	60	\leftarrow	S2	15		
	S3	19		18		73			S2	15		21 18 26	53		53	10		
	S4	21		36		64			S3	19			73		СЛ СЛ	71		
	S5	12		26		56			54 85	21 12		30	64 56		04	<u> </u>		
									Δ	1Z 53		20 59	50	<u> </u>	55	12		
HBase Database									B	50		63	45					
	Name	A e	B	C	D	F	F		C	48		51	63					
			50	10			F 0		D	50		63	70		Name	A		
	POS X	53	50	48	50	42	58		Е	42		49	66		Pos X	53		
	Pos Y	/ 59	63	51	63	49	57		F	58		57	62		Pos Y	59		
	Value	e 59	45	63	70	66	62								Value	59		
												value						
Total Aggregation Time													Tota	al Agg				
												l (st	100.00					
													100,00					
									• • • • • • • • • • •	.1		000	10,00					
									sequentia			(se	1 00)				
10	I	100		1000	ר	100			arallel V		e	1,00	1	Λ	100			
		100		1000)	TUC	000						0,10	I	0	100		
													0.01					
Number of entries per DB																		
	-		R	A		S	iC	S		ne Scotti omputer	sh Info Scienc	rmatics e Alliar	& ice	CO	MM]	T /		
NAL	Augme researc	Center for Ir ented Researcl h • collaborati	nternet h & Asso on • sch	essment nolarship			- O edi data-int - O re	nburgh ensive search	Ci	SC		HIM PH		××××	Un	JIVERS		

		INIYOQL		126		,		INIYOQ						
	ID	Coord X	Coord Y	Intensity							ID	Coord >		
ole	S1	10	25	60		Name	Χ	Υ	Value		S1	10		
	S2	15	21	53		S1	10	25	60	~	S2	15		
	S 3	19	18	73		S2	15	21	53		63	10		
, o ,	S4	21	36	64		S3	19	18	73	-	00 01	18 01		
		12	26	56		S4	21	36	64 50		34			
ea Ilfe			20	20 30		55 ^	1Z 53	20 50	50	<	S5	12		
lra Ira		HBas	e Datab	ase		R	50	63	59 45					
Pa Pa	Nam	Δ	R C		F	C	48	51	63			нва		
e –						D	50	63	70		Name	e A		
Ö	Pos	X 53	50 48 3	50 42	58	E	42	49	66		Pos X	53		
	Pos	Y 59	63 51 6	63 49	57	F	58	57	62		Pos Y	<i>′</i> 59		
	Valu	e 59	45 63	0 66	62							50		
	_										value	59		
00	Tot	al Agg	regatio	n Time					4000 00		Tot	al Agg		
00				_					1000,00					
00								ds	100,00					
00								U U	10.00					
00					-	Sequenti	al	e c	10,00					
00						' Parallel \	/1 0	S S	1,00					
10 100 1000 10000)0		, 1.0	Lime	0,10	1	0	100		
01									0.01	•				
Number of entries per DB											Number of			
ERSITY OF CAGO					CS	a * 3	The Scotti Computer	sh Informatics Science Alliar		CO	[MM]	[T /		
ORIDA TERNATION IVERSITY	NAL Augn resea	Center for Internet Augmented Research & Assessment research • collaboration • scholarship				urgh sive arch				UNIVERS:				









Data Service Functionality:

Aggregate data from different sources (DB providers) according to some semantically meaningful operations.

DB provider <> DS: The databases register their URLs through the DS API. The Data Service creates and maintain a list of registered DB and their

information.

Parallelization and Results

Two different parallelization approaches were applied here to solve the data aggregation process among different database paradigms. Using single threads to aggregate the data in each database led to performance decrease. However, dividing the data inside a single database, then using threads to aggregate it, and then repeating the process for other databases worked much better than the initial approach.

The authors of this work would like to thank the Partnership for International Research (PIRE) program, the Open Science Data Cloud (OSDC), and the National Science Foundation for their support of this work. We would like to extend our special thanks to Dr. Heidi Alvarez, Vasilka Chergarova, and all the members of the Center for Internet and Augmented Research and Assesment (CIARA) at FIU.

Conclusion

Acknowledgments

