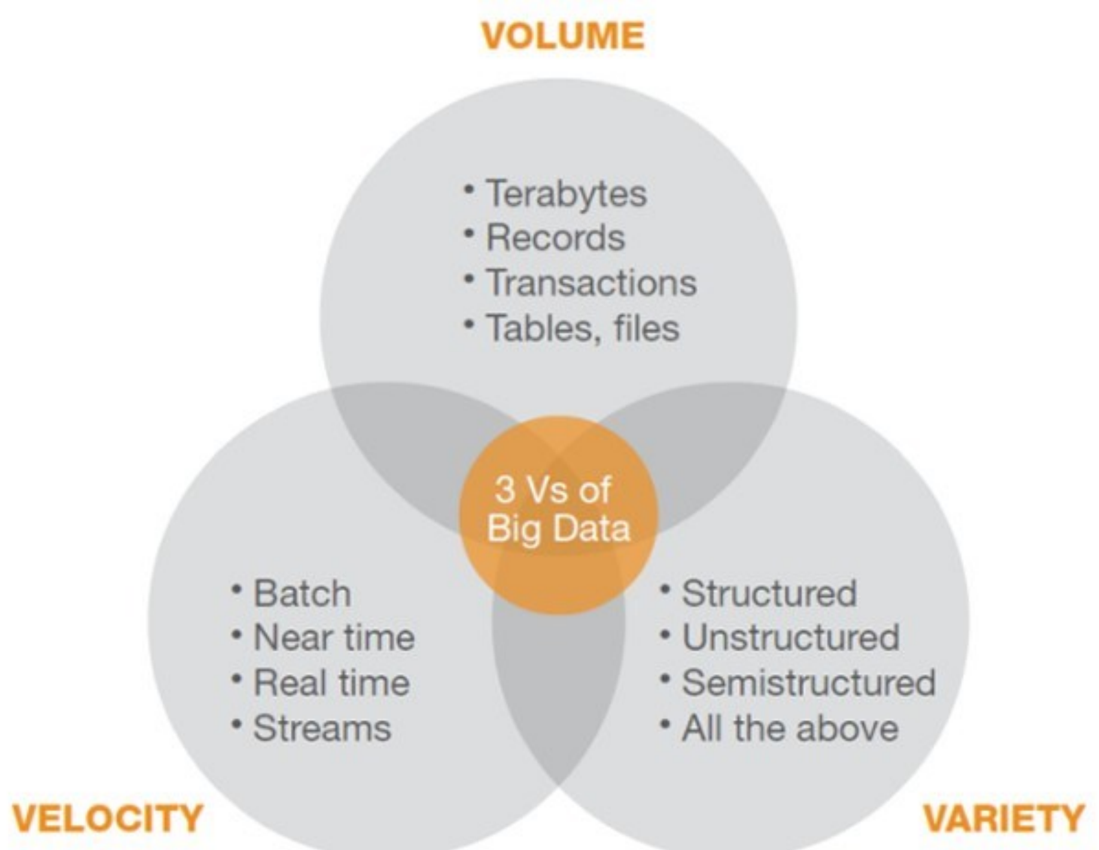
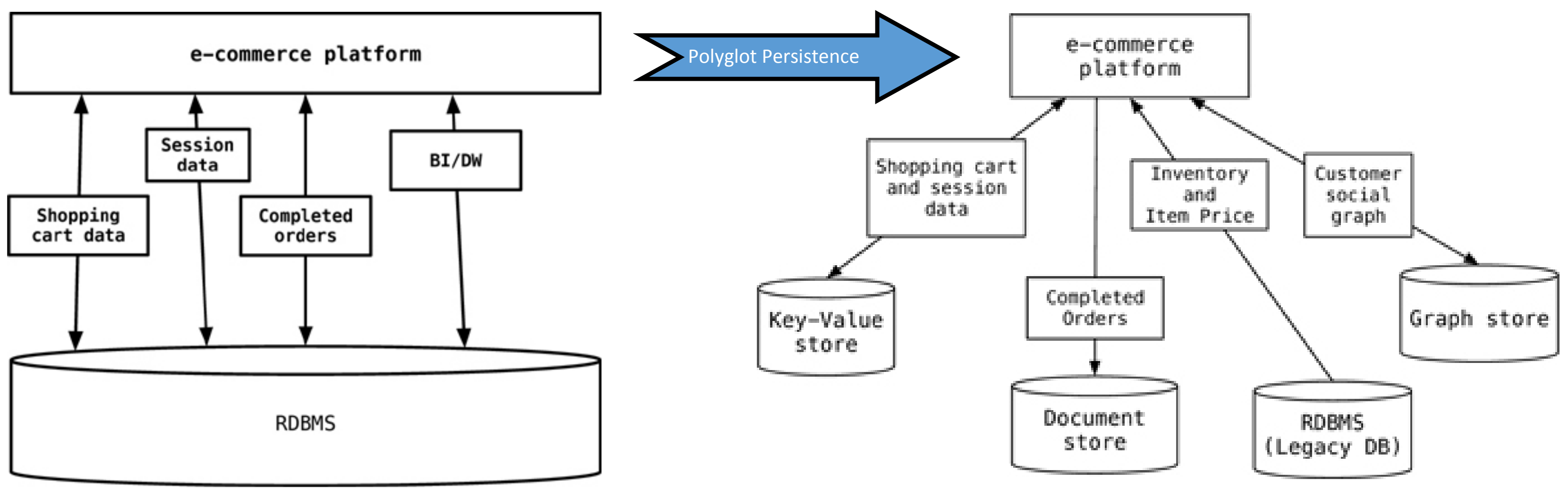
Optimizing Analytic Data Flows in Polyglot Persistence

Rana Faisal Munir, Alberto Abelló, Oscar Romero
Universitat Politècnica de Catalunya, BarcelonaTech
[fmunir | aabello | oromero]@essi.upc.edu

Wolfgang Lehner, Maik Thiele
Technische Universität Dresden
[wolfgang.lehner | maik.thiele]@tu-dresden.de





http://www.informit.com/articles/article.aspx?p=1930511

Challenges

- Perform data analysis in polyglot persistence
- Different query processing capabilities
  - Joins
  - Range queries
  - Secondary Indexes
  - Aggregation
- Optimization of analytic data flows
- Selection of store and storage format for intermediate results

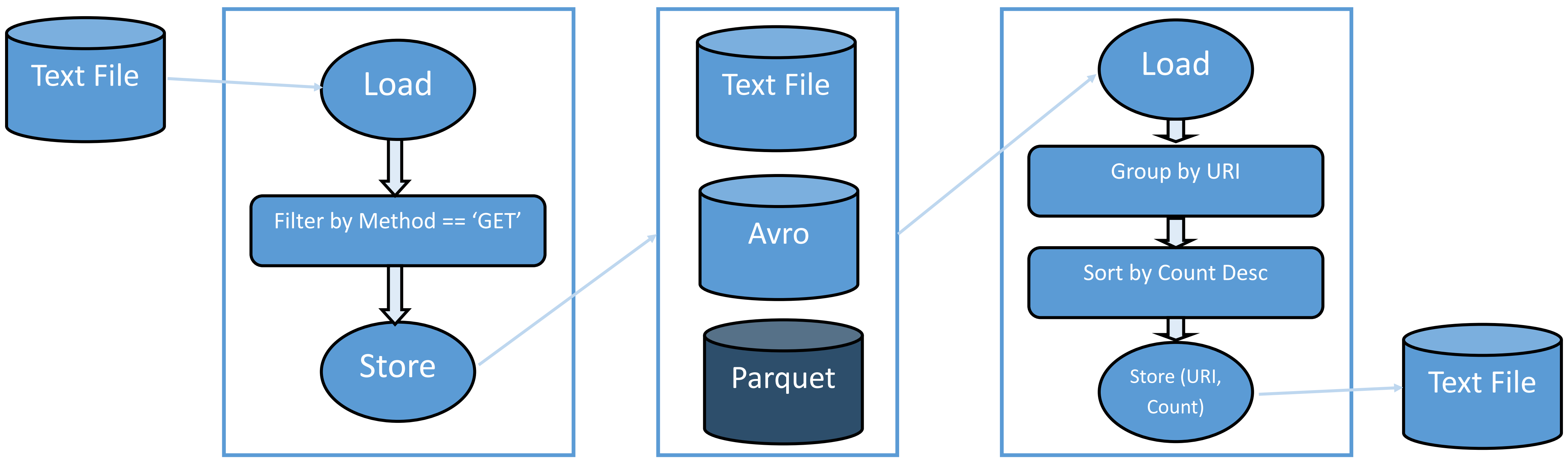
Motivation

- Iterative workflows in real workloads
- 80% of data re-accesses occur on the range of minutes to hours (VLDB 2012)
- Workflows of different users share computation
- Reuse of shared computation save storage and computation cost
- Selection of data store based on access characteristics of intermediate results gives speedup

Our Research Objectives

- Better utilization of the capabilities of data stores
  - Fully utilize the features of data stores
- Reuse of intermediate results
  - Data flows produce intermediate results
  - Results can be materialized for future reuse
- Selection of store for intermediate results
  - Based on access characteristics of intermediate results

Evaluation



Text File			Parquet			Avro		
Start	End	Total	Start	End	Total	Start	End	Total
11:12:15	11:18:02	00:05:47	11:18:32	11:23:36	00:05:04	12:44:17	12:50:00	00:05:43
12:13:40	12:19:03	00:05:23	11:26:01	11:30:57	00:04:56	12:55:58	13:01:32	00:05:34
12:19:30	12:24:57	00:05:27	11:42:36	11:47:36	00:05:00	13:04:24	13:09:48	00:05:24
12:25:20	12:30:46	00:05:26	11:58:37	12:03:35	00:04:48	13:13:15	13:18:48	00:05:33
12:31:29	12:37:09	00:05:40	12:05:24	12:10:27	00:05:03	13:29:30	13:35:19	00:05:49
	Average	00:05:23		**Average**	00:05:00		**Average**	00:05:37

Existing Frameworks with Polyglot Persistence Support

Features	Apache Spark	Apache Drill	SQL++
Data Structure	RDD	JSON	Similar to JSON
Query Language	SparkSQL	Similar to ANSI SQL	Similar to SQL
Architecture	Connectors for different data stores	Connectors for different data stores	Mediate-Wrapper
Query Optimization	Apache Catalyst	Apache Calcite	No
Local Query Optimization	Predicates Push Down and Column Pruning	Predicates Push Down and Column Pruning	No
Scalable	Yes	Yes	No
Aggregation Framework	Yes	Yes	No
Data Locality	Yes	Yes	No
Company	Databricks	MapR	UC San Diego, USA

Barcelona
Information Technologies for Business Intelligence Doctoral College
Erasmus Mundus Joint Doctorate
IT4BI DC

Fifth European Business Intelligence Summer School (eBISS 2015), Barcelona, Spain