

# SEDA: A SYSTEM FOR SEARCH, EXPLORATION, DISCOVERY AND ANALYSIS OF XML DATA

Andrey Balmin<sup>†</sup> Latha Colby<sup>†</sup> Emiran Curtmola<sup>\*</sup> Quanzhong Li<sup>†</sup> Fatma Özcan<sup>†</sup> Sharath Srinivas<sup>Δ</sup> Zografoula Vagena<sup>‡</sup>  
<sup>†</sup>IBM Almaden Research Center <sup>\*</sup>UC San Diego <sup>Δ</sup>UMD <sup>‡</sup>Microsoft Research

## • Problem

– Extracting meaningful insights from heterogeneous XML data collections  
 e.g. Find the average **trade percentage** amounts for **import partners of United States**

## • Challenges

– No fixed schema (driven by schema integration and evolution)  
 – Need to query both XML structure and text, and do not exactly know the structural constraints  
 – Need to compute analytics from semi-structured data (e.g., avg, min, max, count)  
 – Need 100% precision and recall for meaningful analytics, but keyword query results are imprecise

## SEDA

### Keyword Search

- Maximum flexibility
- Queries are underspecified: hard to capture users' intentions
- Results are imprecise and ranked: inaccurate aggregates

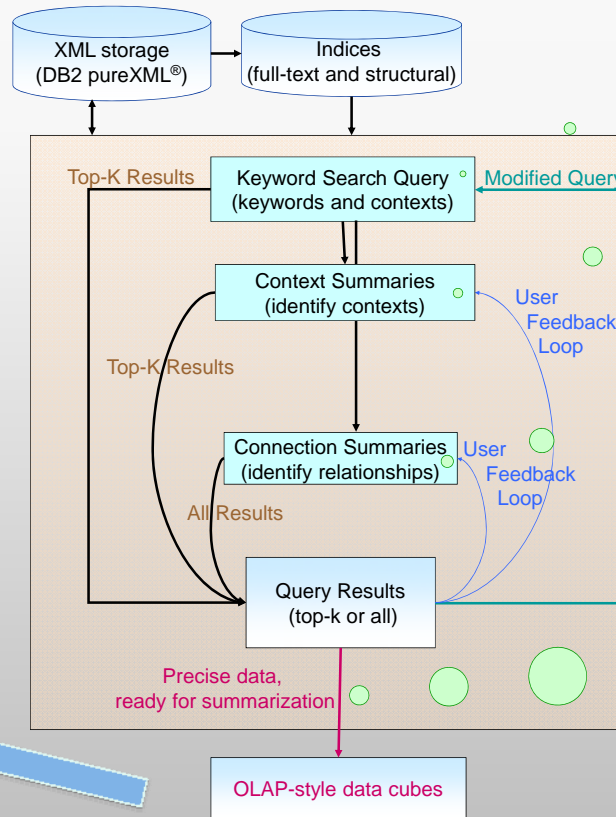
**New Paradigm: Start with simple keyword search, employ user guidance to compute complex OLAP-style analytics**

- Provide quick answers through *ranked retrieval* that
  - Return an initial set of possible answers
- Enable effective *user interaction* via user feedback loops
  - Runtime discovery of XML contexts and connections between nodes
- Use relational data cube semantics to compute summarizations

### XQuery, SQL, SQL/XML

- Complex, hard to express
- Queries are well-specified, and have clean semantics
- Results are exact: accurate aggregates

## SEDA System Overview



### Input query terms

- Context (tag name, path) and search term (Keywords)

### Context Summary (List of Paths)

Different paths may correspond to different real-world entities  
 → Let the user disambiguate paths

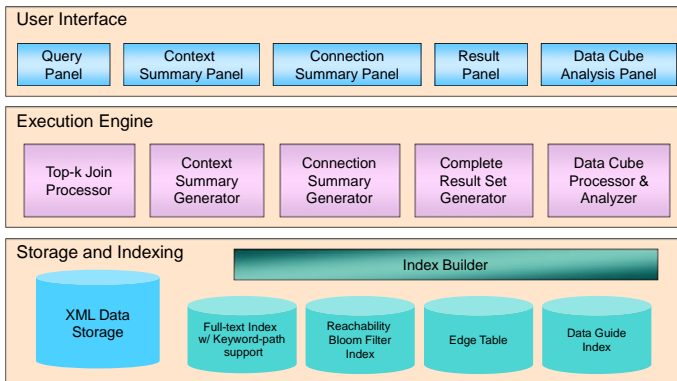
### Connection Summary

- Different relationships
  - Let the user choose the meaningful connections between the query terms
- Infeasible to show all connections
- Compute and show only connections in top-K results, and exploit context filtering

### Data Cube Computation

- Output is a table, one column for each query term,
- Consider query result as a de-normalized fact table
- Match each column in the result to a known dimension or measure
- Augment the query results with keys and values, as needed
- Compute the normalized dimension and fact tables
- Feed result into DB2 AlphaBlox® to compute aggregations and investigate

SEDA System Architecture



### List of Dimensions

Name	Context	Key
trade-country	/country/economy/import_partners/item/trade_country	(/country,/country/year,)
country	/country	(/country,/country/year)

### Query result (de-normalized fact table)

id1	path1	id2	path2	id3	path3
n1	/country	n2	/country/economy/import_partners	n3	/country/economy/import_partners/item/percentage
n4	/country	n5	/country/economy/import_partners	n6	/country/economy/import_partners/item/percentage

### Fact Table (for percentage)

Country	Year	Partner Country	Trade Amount
United States	2004	China	12.5
United States	2004	Mexico	10.7
United States	2005	China	13.8
United States	2005	Mexico	10.3
United States	2006	China	15
United States	2006	China	16.9
...	...	...	...
United States	2007	China	4.6
United States	2007	Mexico	13.3

### List of Facts

Name	Context	Key
import-trade-percent	/country/economy/import_partners/item/percentage	(/country,/country/year, /trade_country)
GDP	/country/economy/GDP	(/country,/country/year)