Schema–Independent Scientific Data Cataloging Framework

Supun Nakandala, Sachith Dhanushka Withana, Dinu Kumarasiri, Hirantha Jayawardena and H.M.N. Dilum Bandara

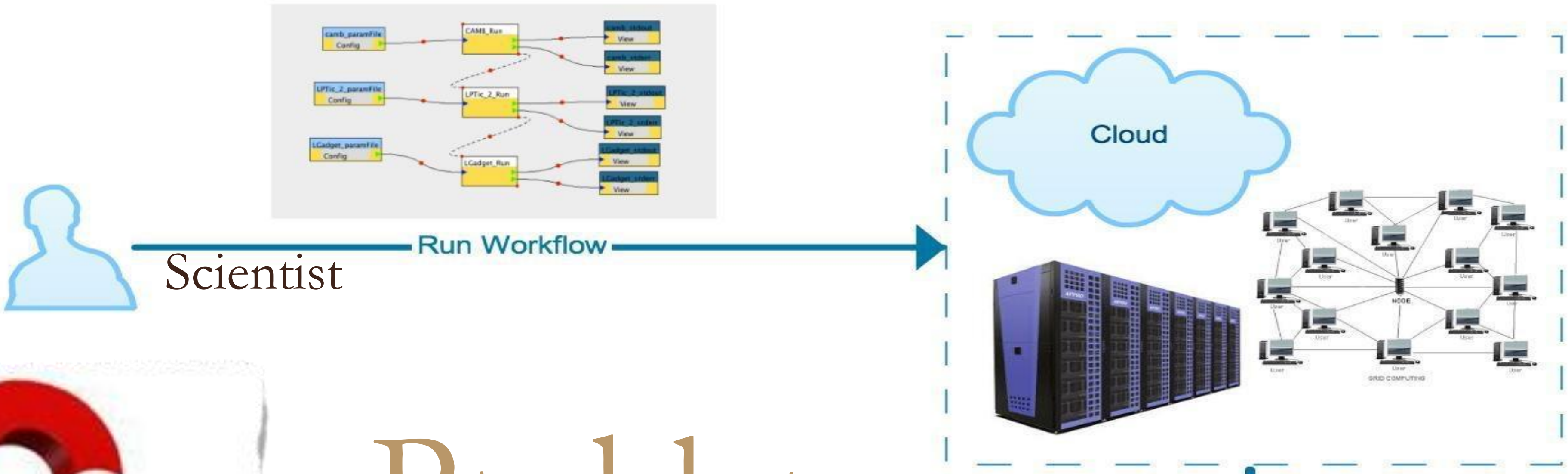
(Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka)

Srinath Perera

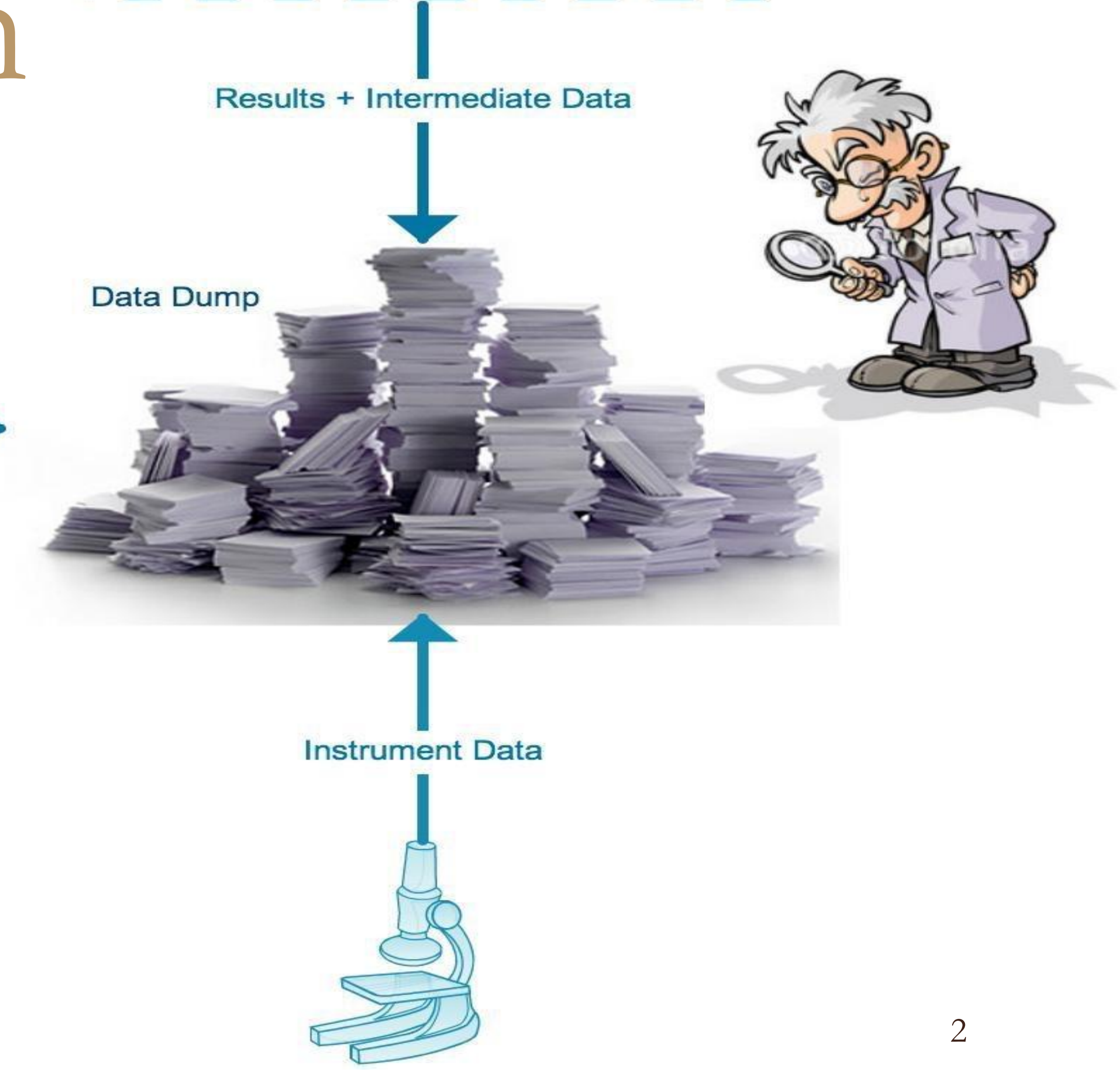
(Wso2 Inc., Colombo, Sri Lanka)

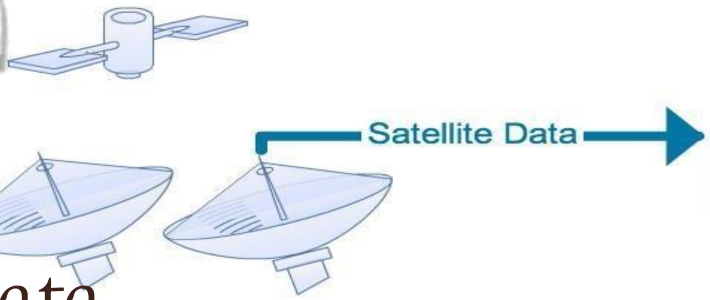
Suresh Marru, Sudhakar Pamidighantam

Indiana University, Bloomington, USA



Problem





Scientific Data

- ✧ Vast Volume
- ✧ Hard to …
  - ❑ Search
  - ❑ Reuse
  - ❑ Share findings

GridChem Usecase

- Gaussian 9 experiments generate vast amount of data in two forms
  - Output file (*.out)
  - Check point file (*.chk)
- Provide efficient searching among these data

Why we need a new one ??

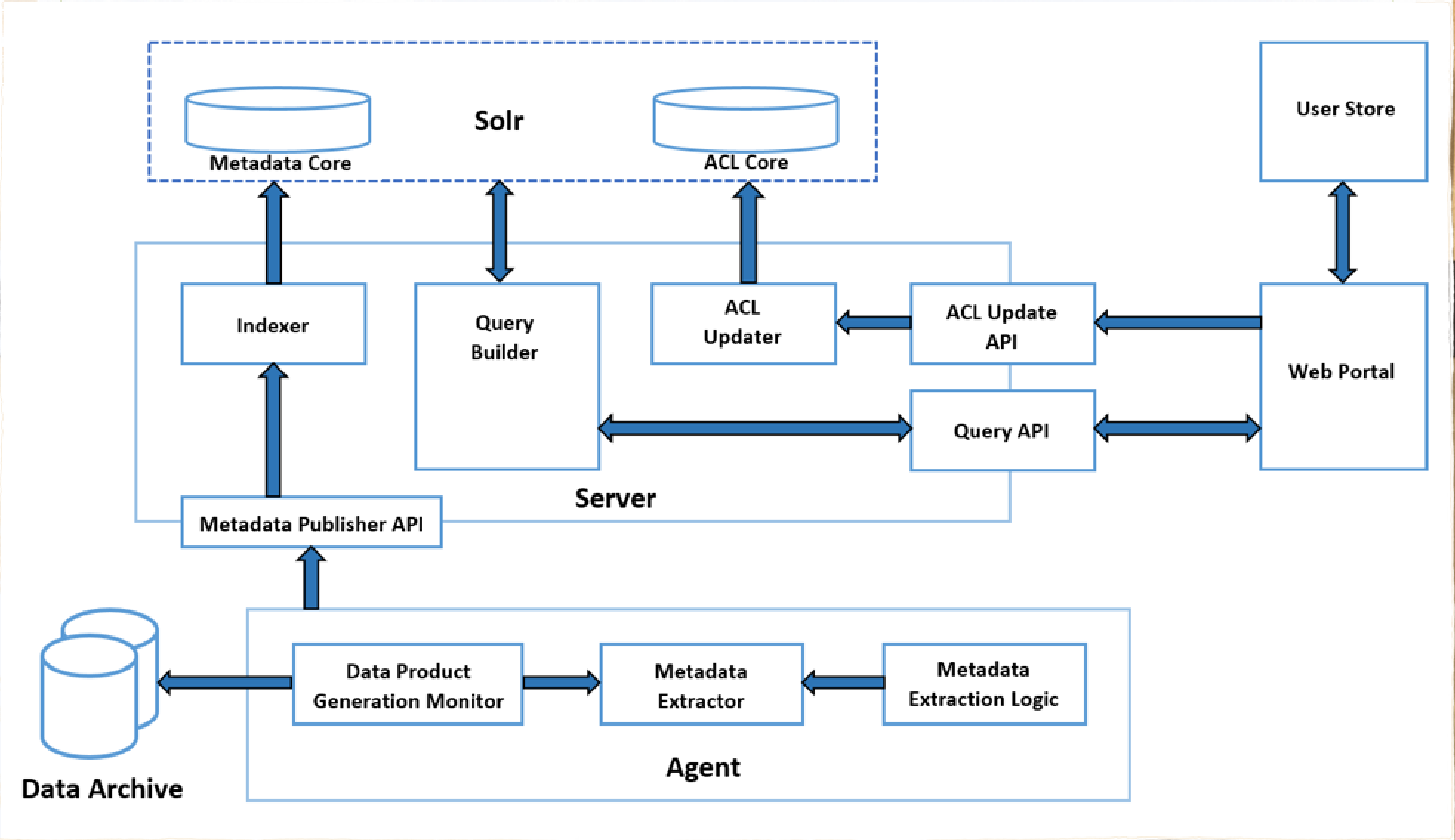
Existing Solutions

- Tightly coupled
- Inflexible querying
- Static schemas
- Eg:-
  - MCS
  - MCAT
  - MyLEAD

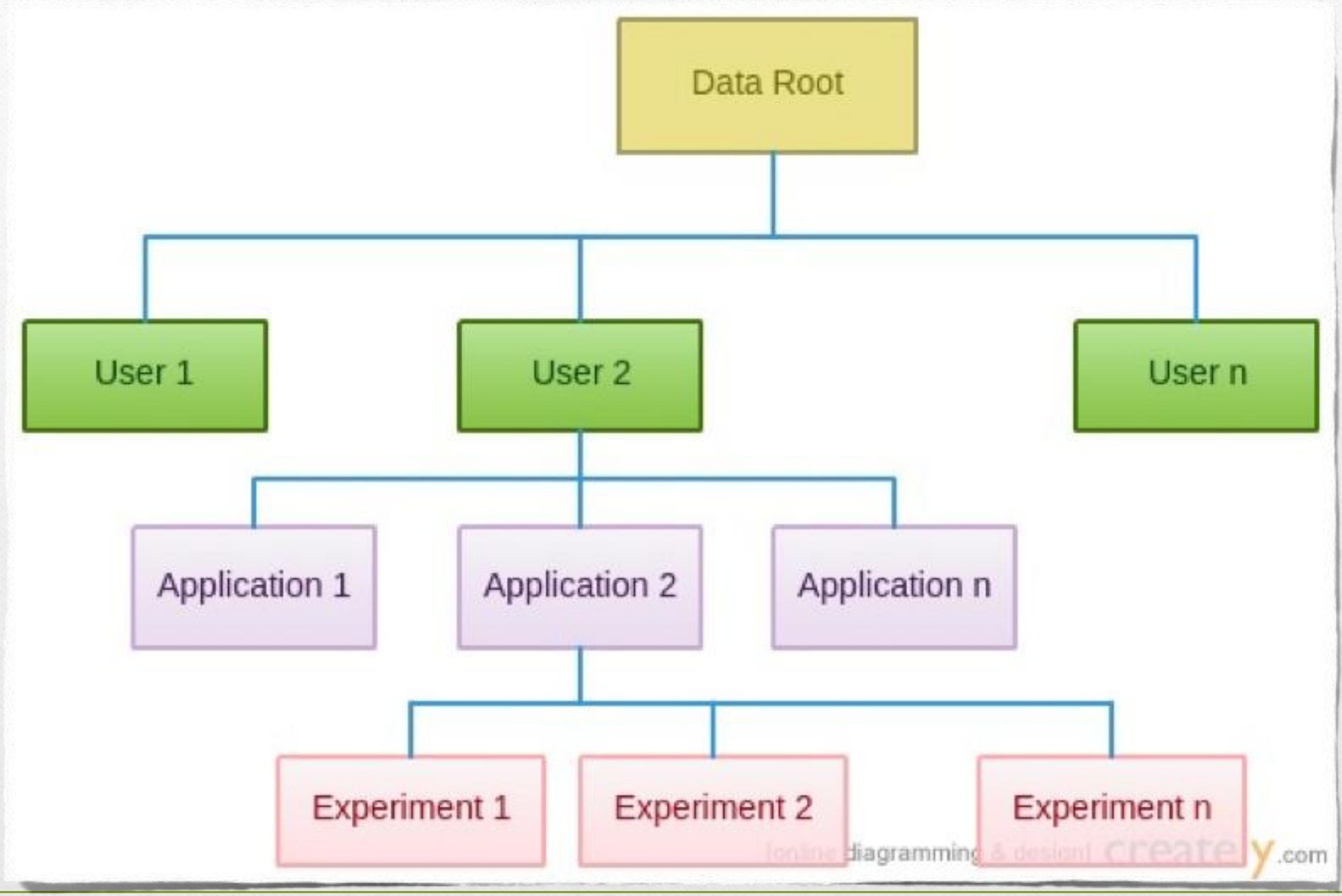
Our Solution

- Generalizable framework
- Flexible querying
  - Wild card queries
  - Full text queries
  - Substring queries
  - Fielded queries
- Static schema + dynamic fields

High-level Architecture



Folder Structure



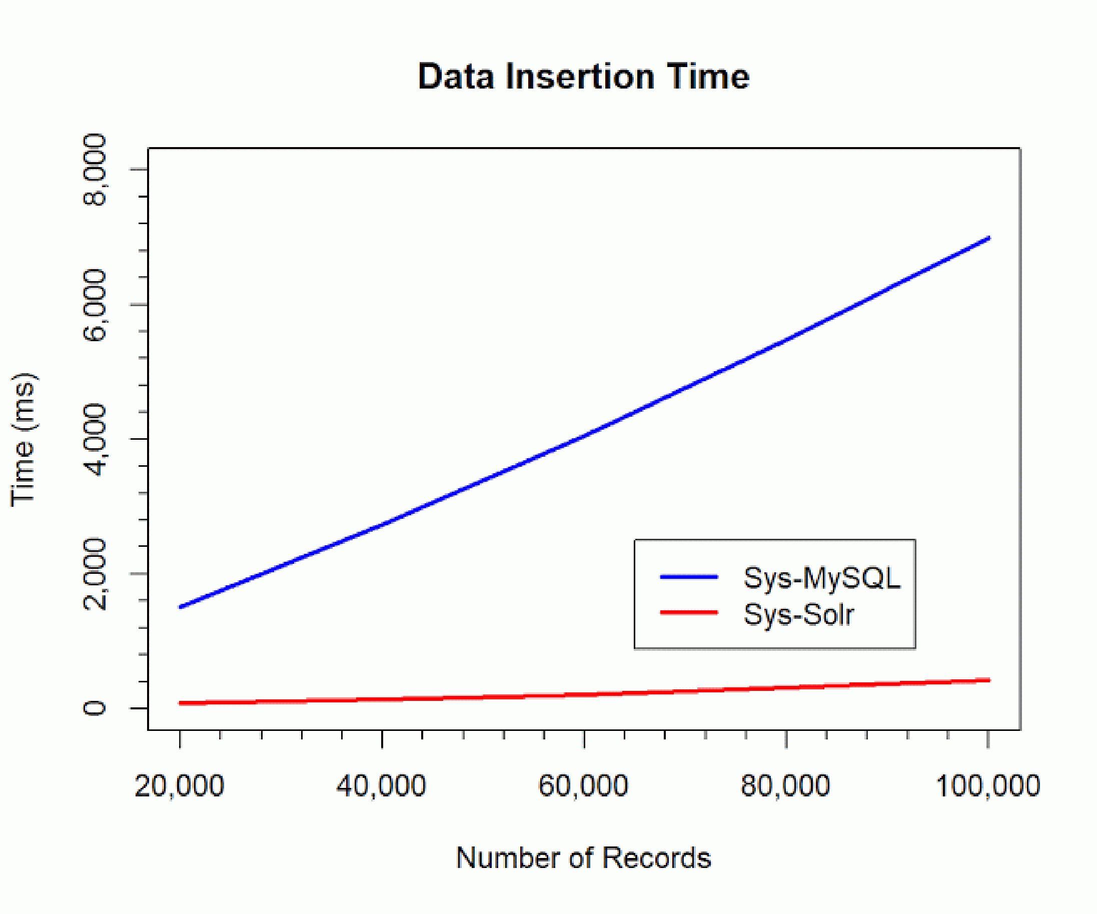
What is new in our solution?

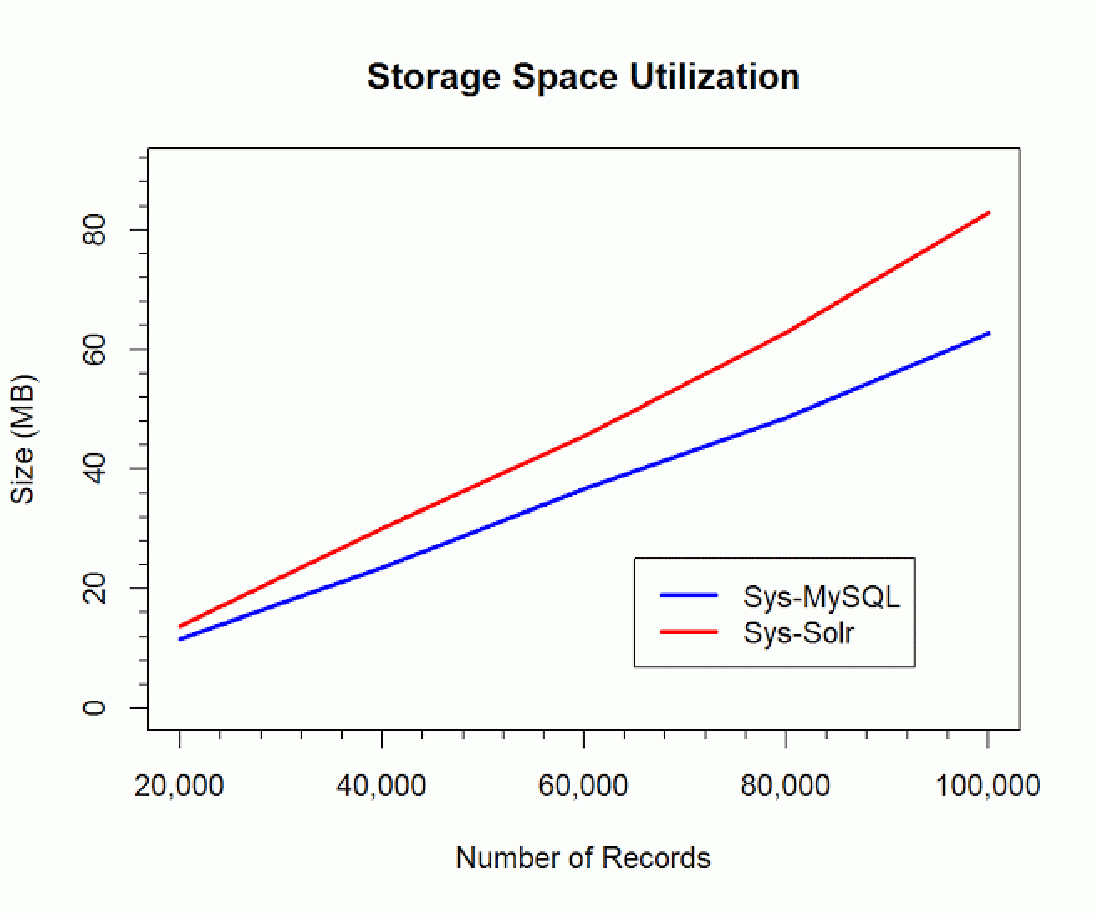
- Pluggable metadata extraction logic
- Extensible data product generation monitors
- Use of NoSQL database (Apache Solr)
- Ability to dynamically add metadata fields

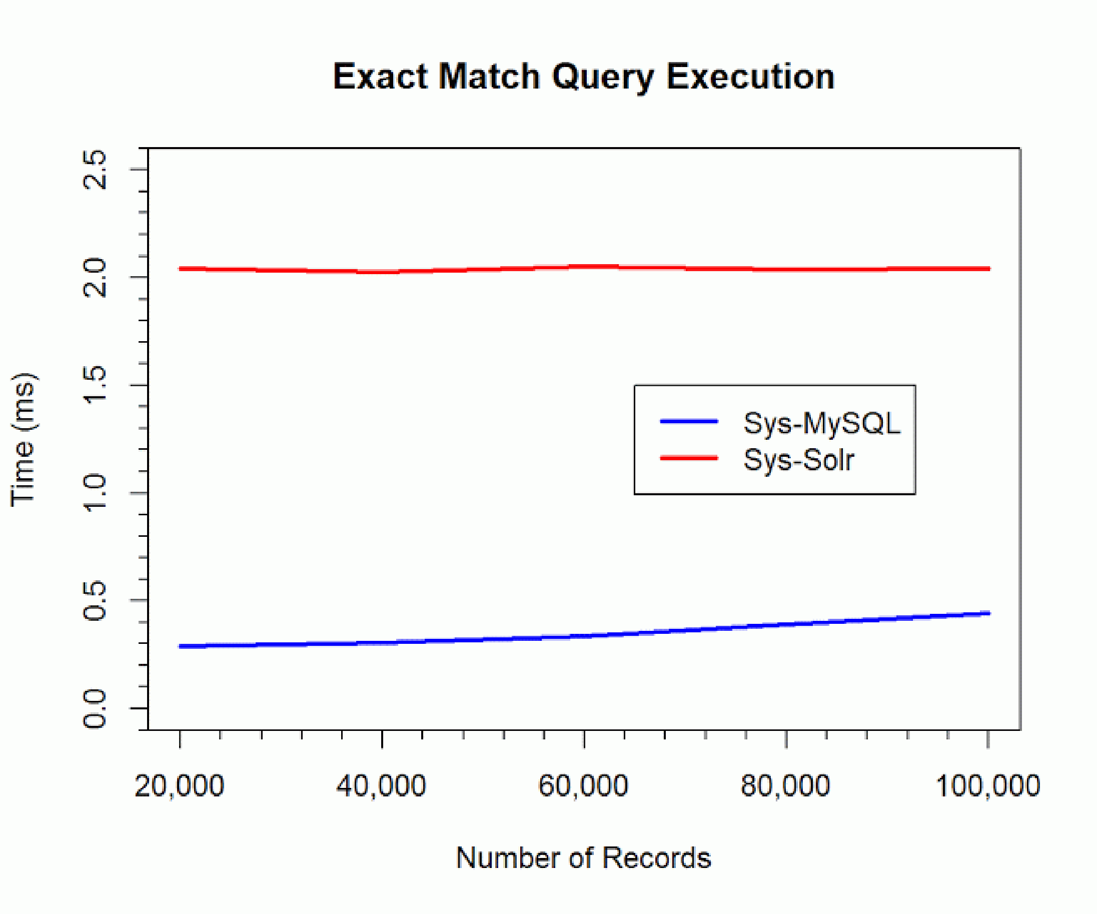
Performance Test

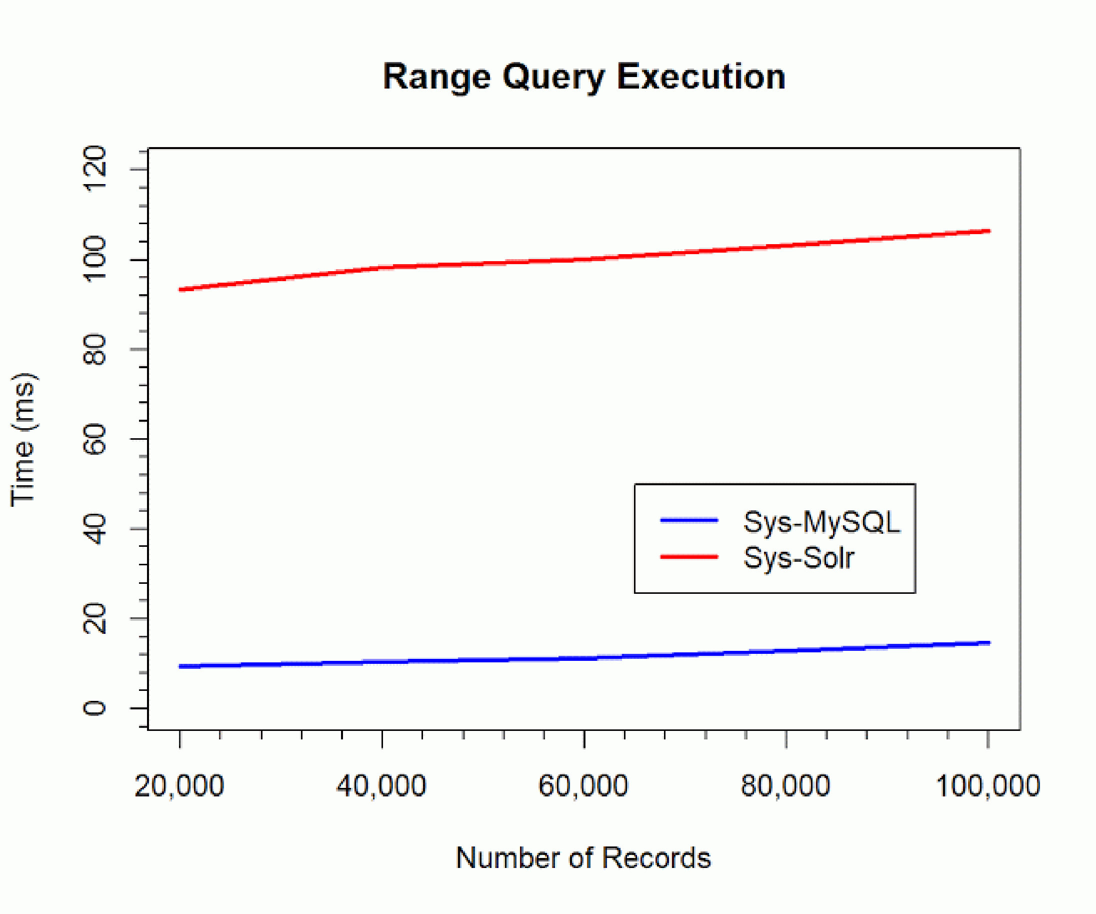
- MySQL vs Solr
- Data Insert Performance
- Query Performance
    - Exact match queries
    - Range queries
    - Full text queries
    - Prefix match queries
    - Suffix match queries
    - Wildcard queries
    - Substring queries

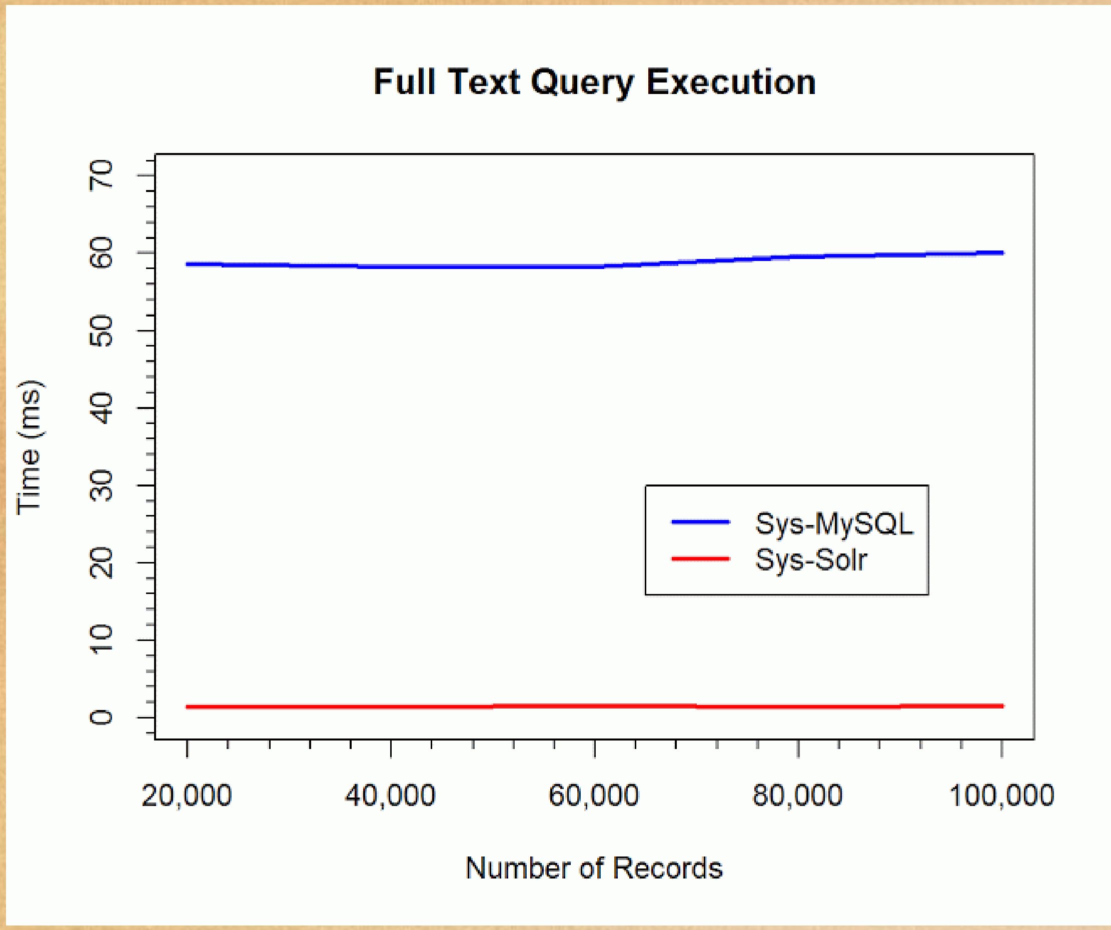
Solr resolves more complex queries 91% - 99% faster than a MySQL-based implementation.

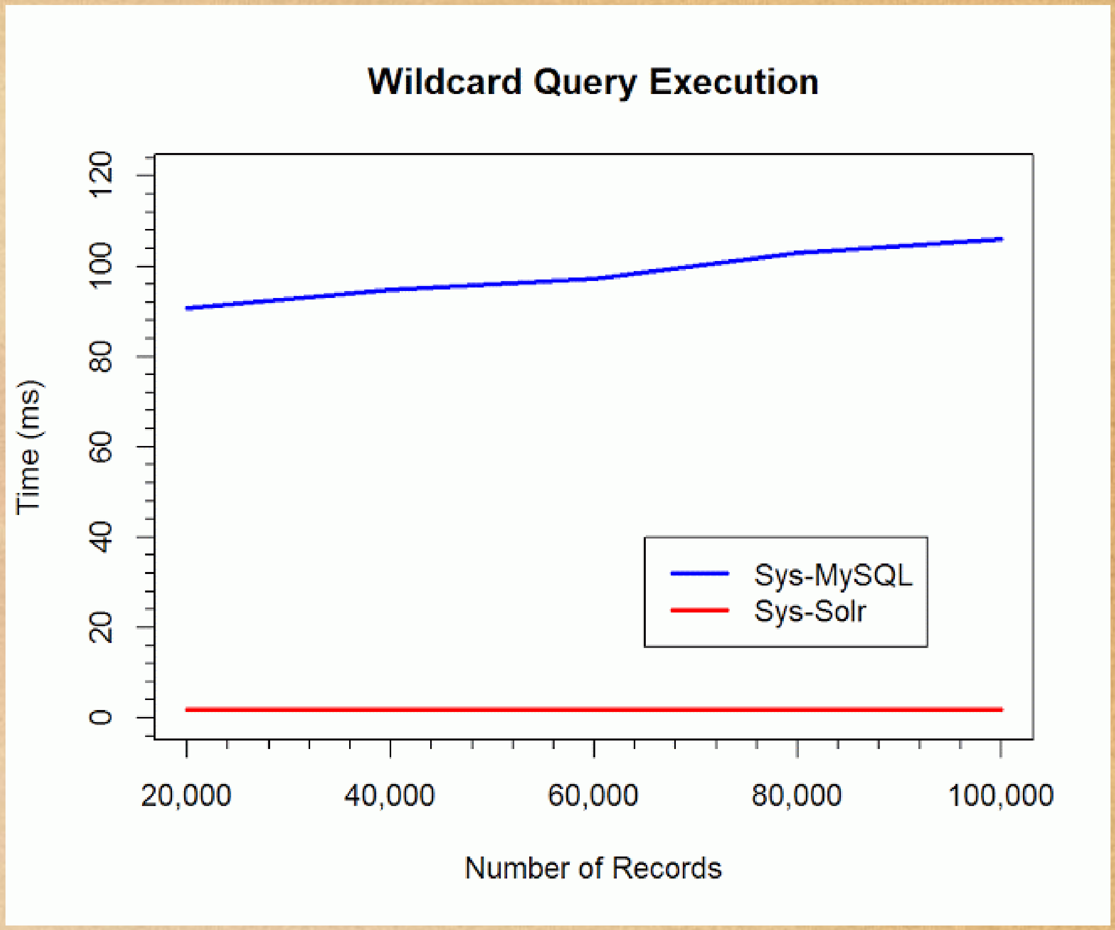

Data Insertion Time
Time (ms)
0
2,000
4,000
6,000
8,000
20,000
40,000
60,000
80,000
100,000
Number of Records
Sys-MySQL
Sys-Solr

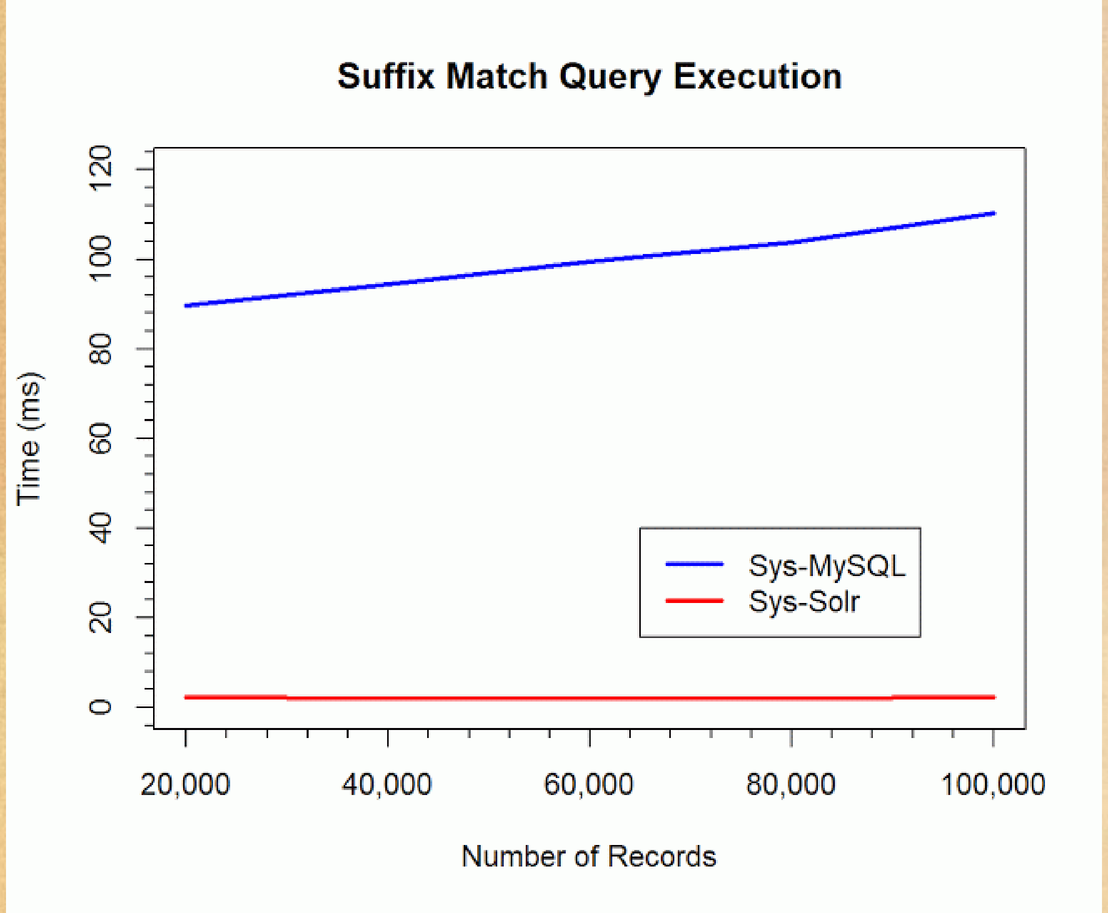

Storage Space Utilization
Size (MB)
0
20
40
60
80
20,000
40,000
60,000
80,000
100,000
Number of Records
Sys-MySQL
Sys-Solr

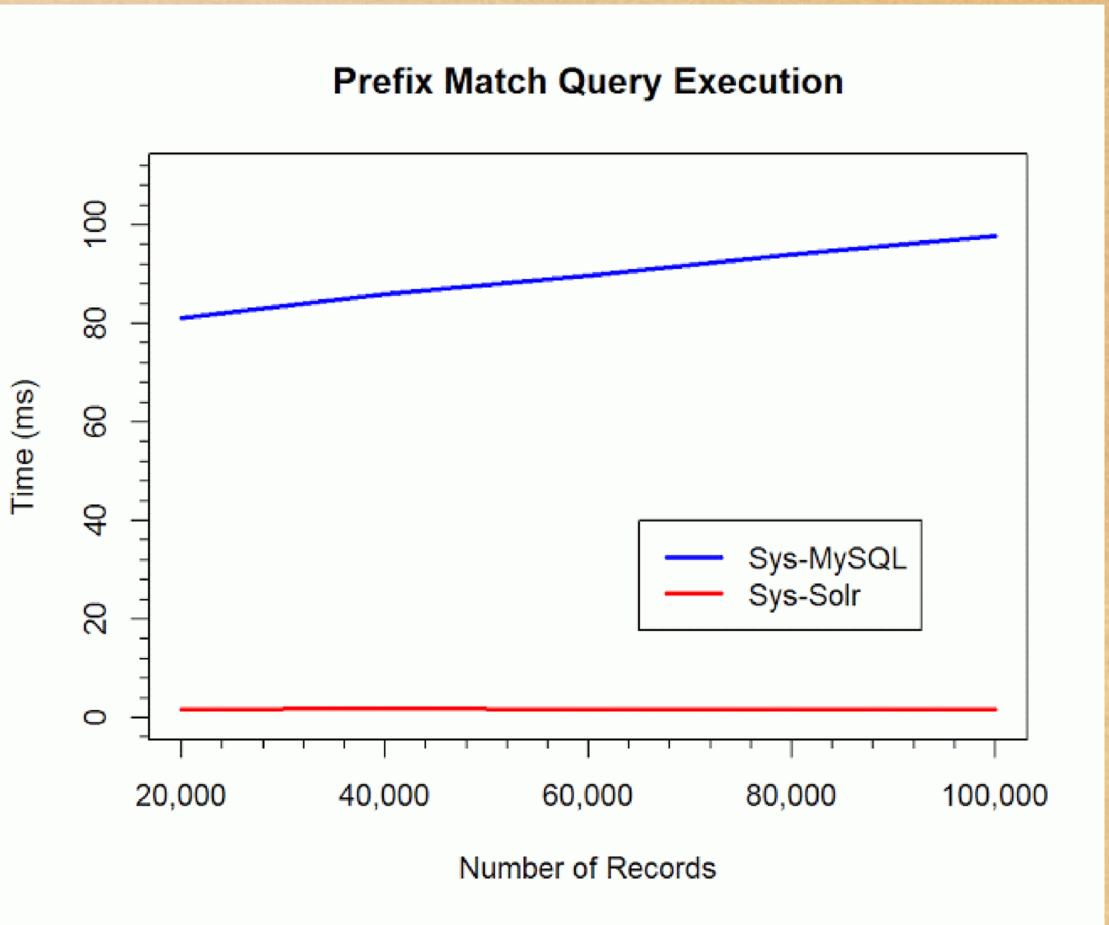

Exact Match Query Execution
Time (ms)
0.0
0.5
1.0
1.5
2.0
2.5
20,000
40,000
60,000
80,000
100,000
Number of Records
Sys-MySQL
Sys-Solr

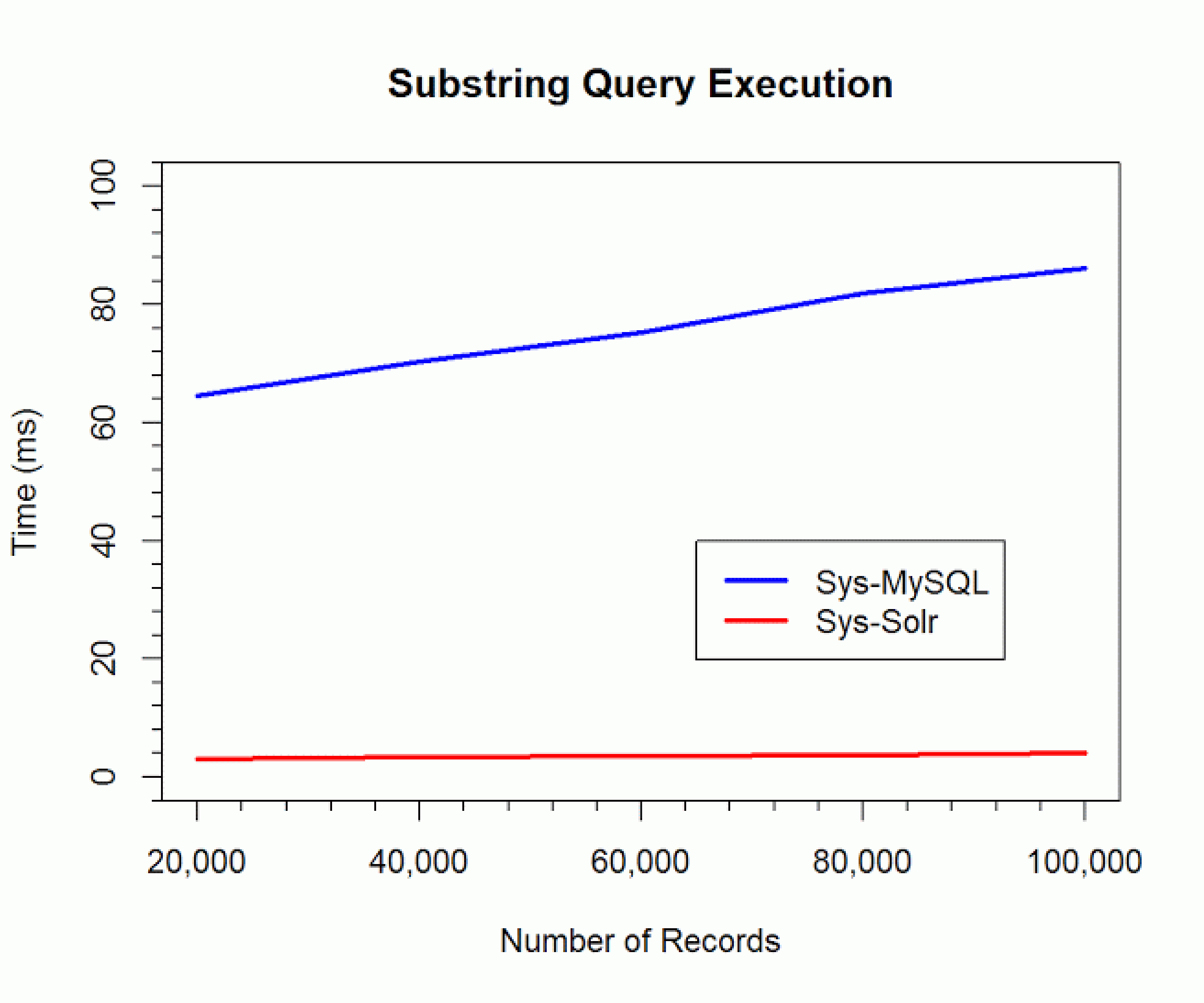

Range Query Execution
Time (ms)
0
20
40
60
80
100
120
20,000
40,000
60,000
80,000
100,000
Number of Records
Sys-MySQL
Sys-Solr


Full Text Query Execution


Wildcard Query Execution


Suffix Match Query Execution


Prefix Match Query Execution


Substring Query Execution
Time (ms)
0
20
40
60
80
100
20,000
40,000
60,000
80,000
100,000
Number of Records
Sys-MySQL
Sys-Solr

Summary

- What we did: A schema-independent scientific data catalog with pluggable parser logic and Solr backend
- Future work: Airavata integration and provenance aware execution

Thank You …