Knowledge Cubes

A Proposal for Scalable and Semantically-Guided Management of Big Data

Amgad Madkour, Walid G. Aref, * Saleh Basalamah

Purdue University, USA
* Umm Al-Qura University, KSA

Motivation

- Understand the **query intent**
  - **Query:** “Michael Jordan Bio”
    - Athlete (Basketball,Baseball) ? Professor (EECS Berkley) ?
      - Understanding the semantics of the the name and Bio

- Utilize **heterogeneous** sources to answer complex queries
  - **Query:** “Michael Jordan Bio”
    - Web ? Encyclopedia ? Social Media ? Most Accurate Source ?

- Architecture that **scales** well to accommodate Big Data sources

Vision

Building systems that are guided by the data semantics that includes topical,contextual, spatial and temporal aspects

- ***Query:*** *"Michael Jordan Bio"*
  - *University campus → Spatial*
  - *Statistics building → Contextual*
  - *"Michael Jordan", "Bio" → Topic*
  - *Recently updated "Bio" → Temporal*

Knowledge Cubes

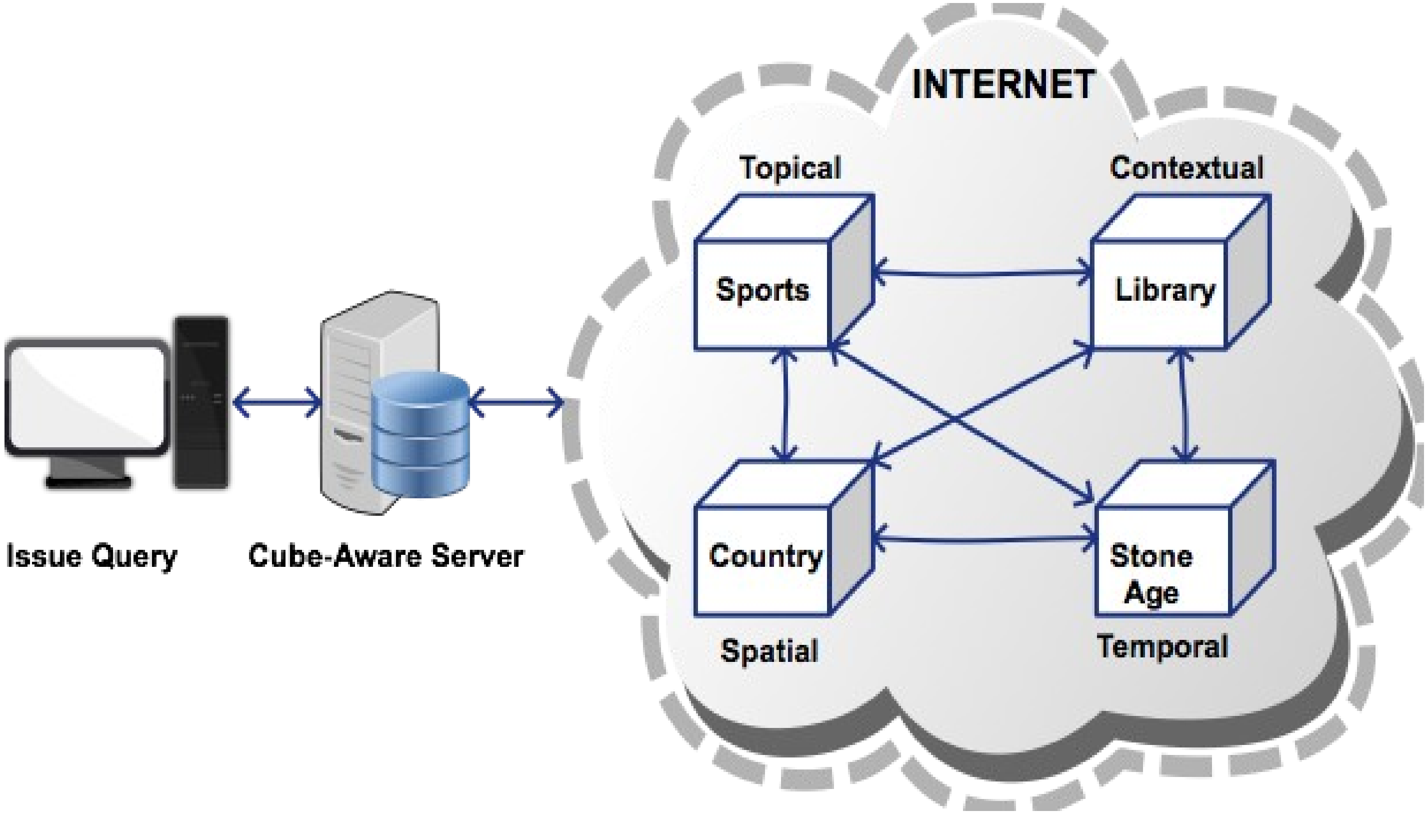
- A database instance capable of storing, analyzing, and searching data
  - ***Intelligent*** → Ingests data and presents accurate answers
  - ***Adaptive*** → Structurally evolves over time

- Established based on semantic aspects:
  - *Topical*,*Contextual*,*Spatial*, or *Temporal*

- ***Specializes*** in handling data only relevant to its semantics

- Uses ***Linked Data*** as its main building block with ***RDF*** as its data model
  - All data in <**Subject, Predicate, Object**> format

Knowledge Cubes

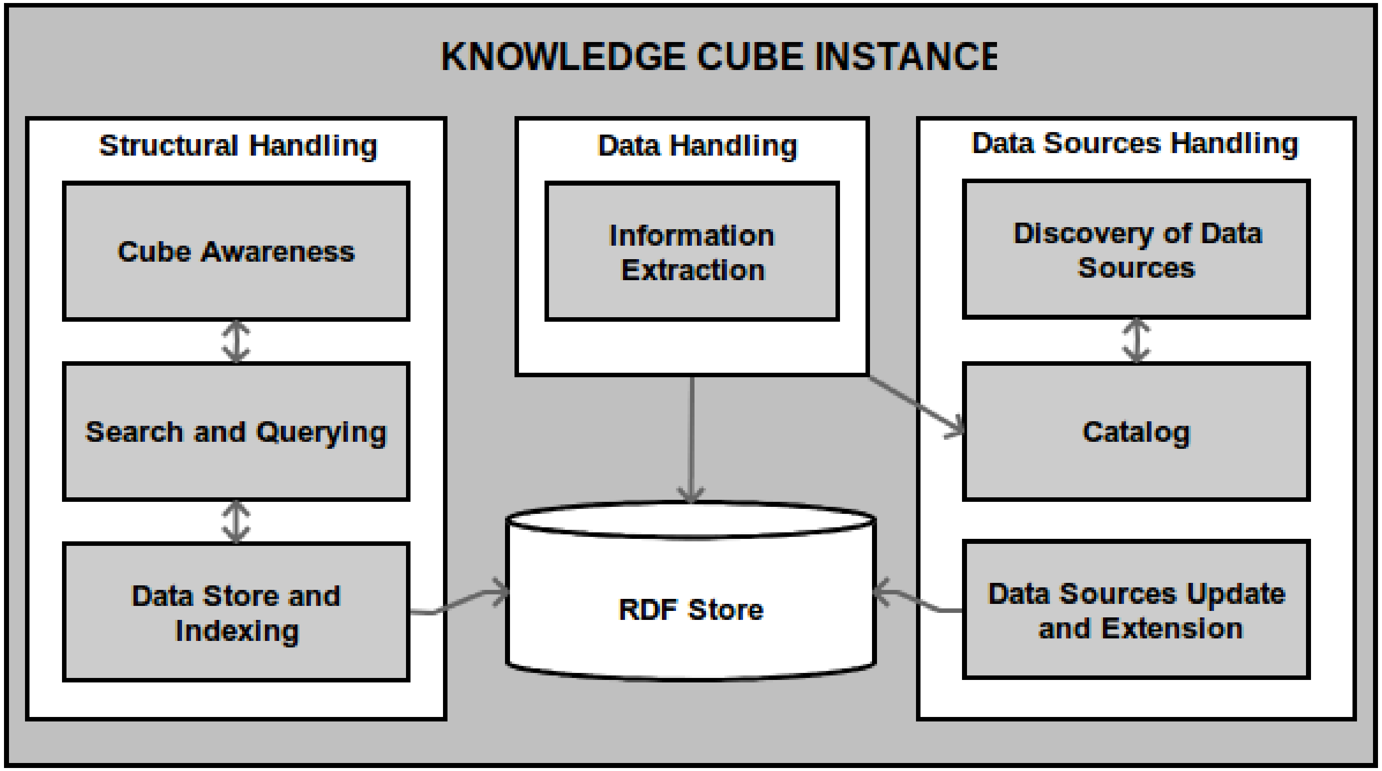


Architecture of Knowledge Cubes

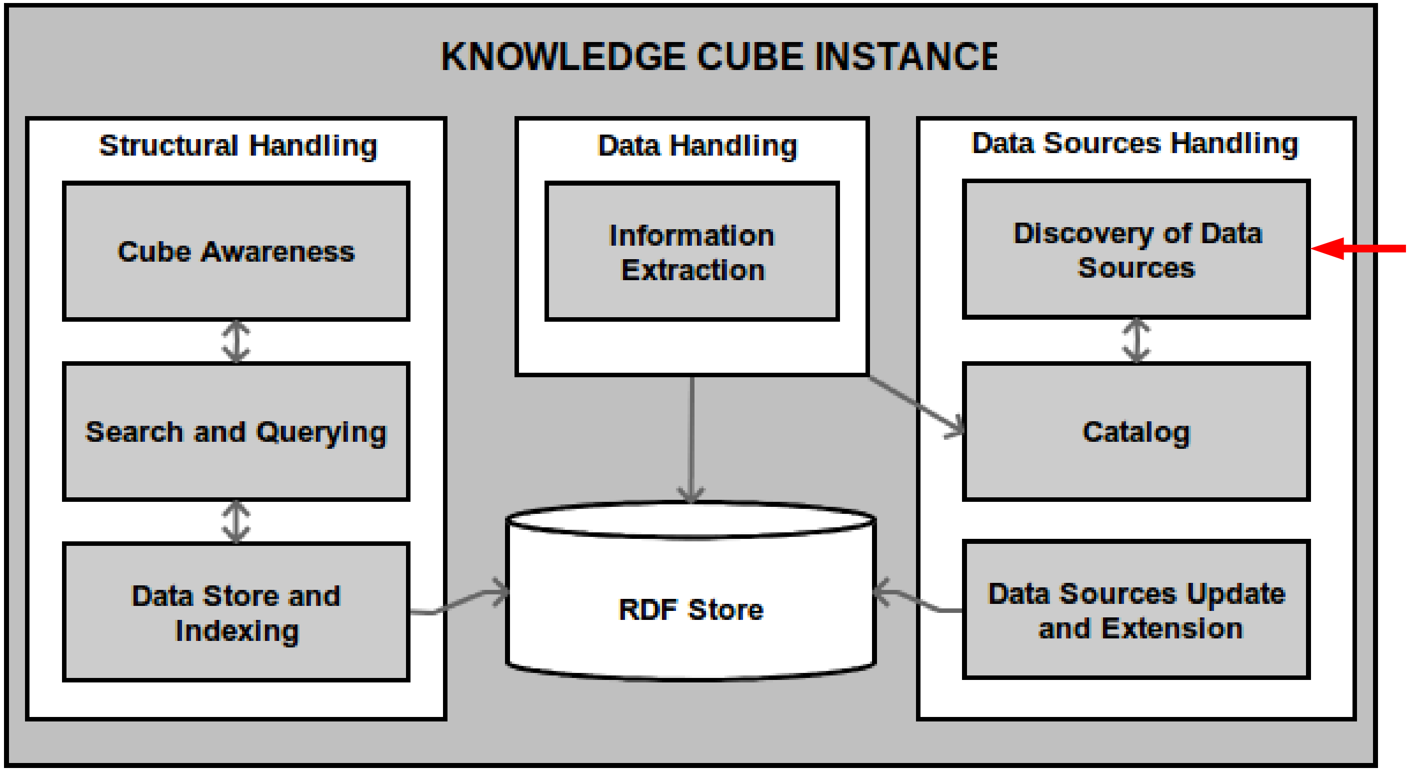
Founding Principles

- **Structural Evolution**
  - Evolves based on its newly attained size or semantic aspect by re-partitioning dynamically in an unsupervised fashion
- **Temporal Evolution**
  - Organizes it own data temporally using a time-roadmap
- **Analytic Distribution**
  - Distributes analytic load across multiple knowledge cubes and then communicates the results back according to relevance

Architecture

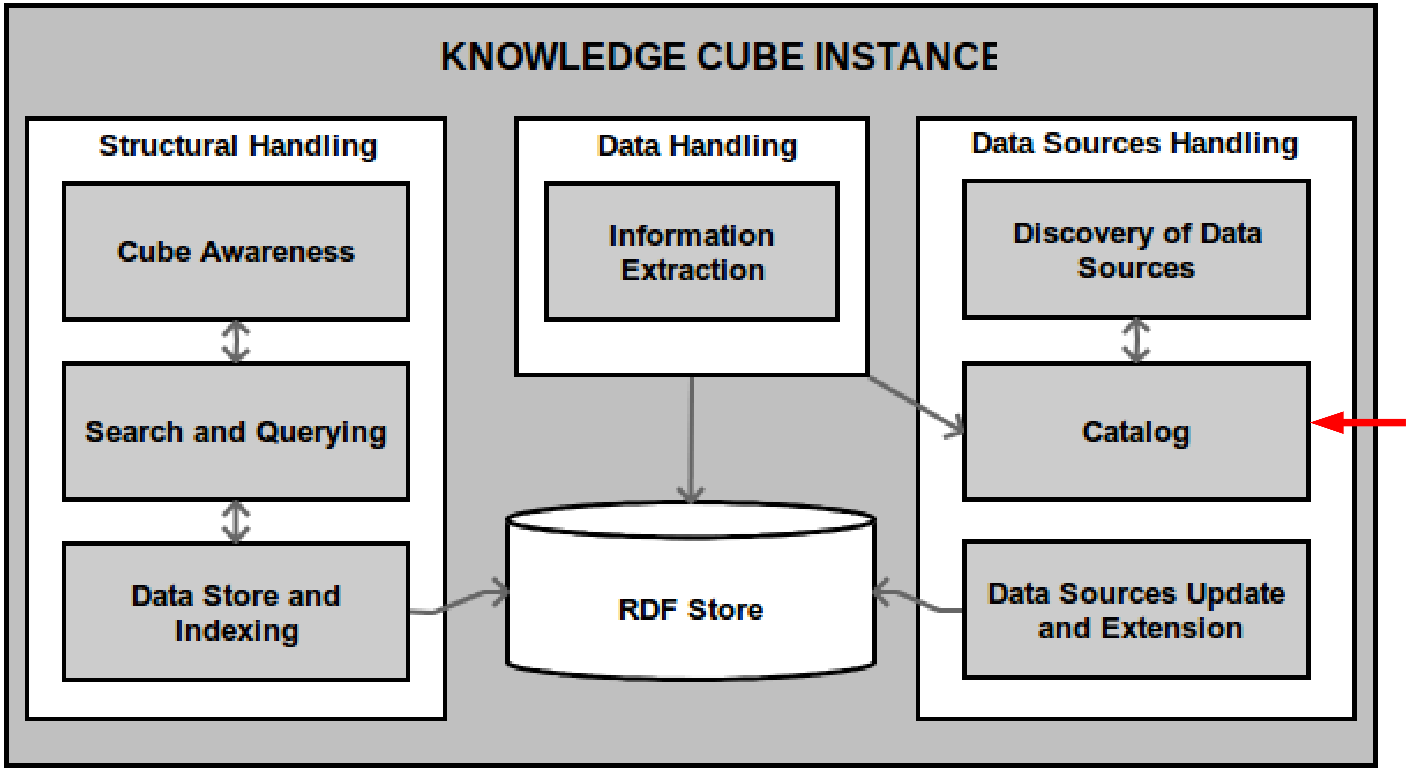


Architecture



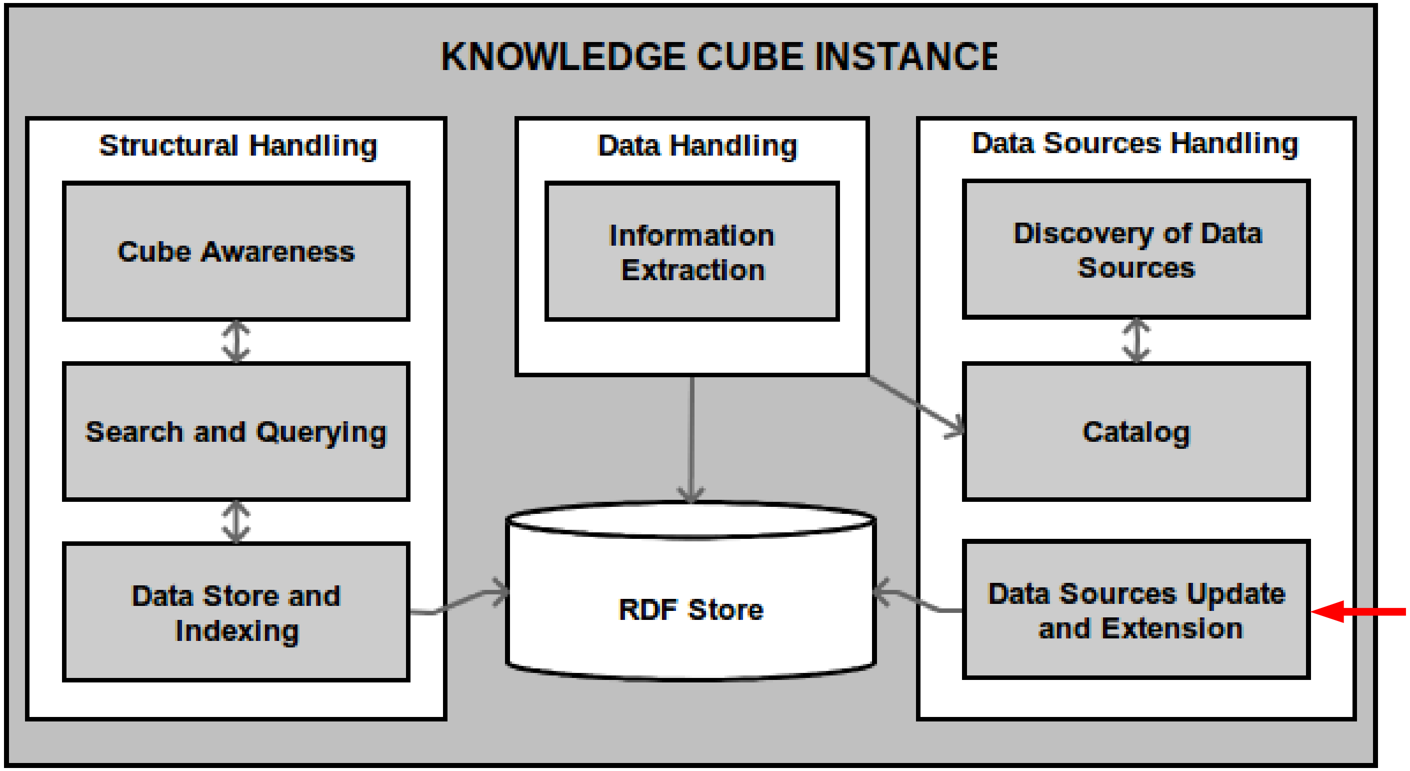
- **Discovery of Data Sources**
  - Maintain and create relationships among data sources
    - **Ex:** Probe the web for relevant data web sources and link them based on their semantics

Architecture



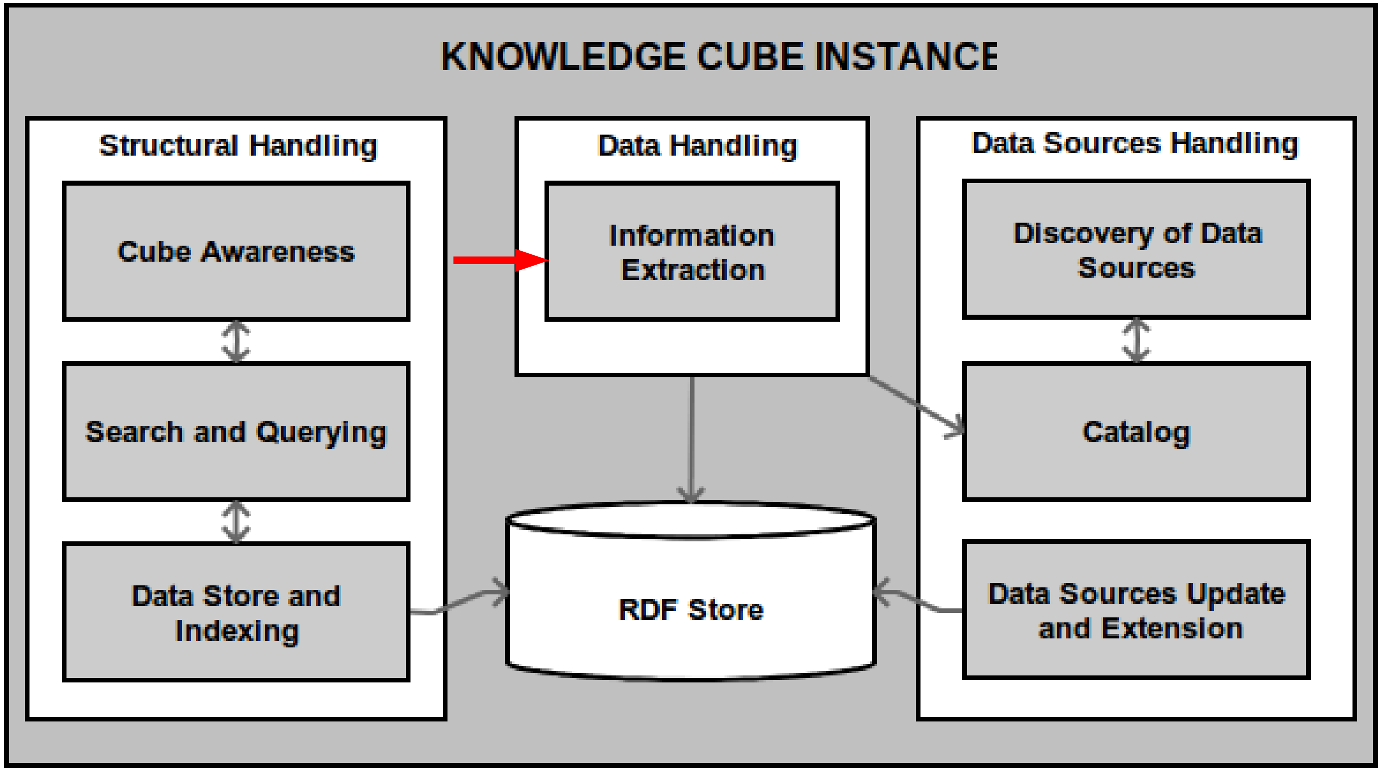
- **Catalog**
  - Maintains all information related to data sources
    - **Ex:** A catalog entry might include Wikipedia so we maintain its meta-information (last-updated etc.)

Architecture



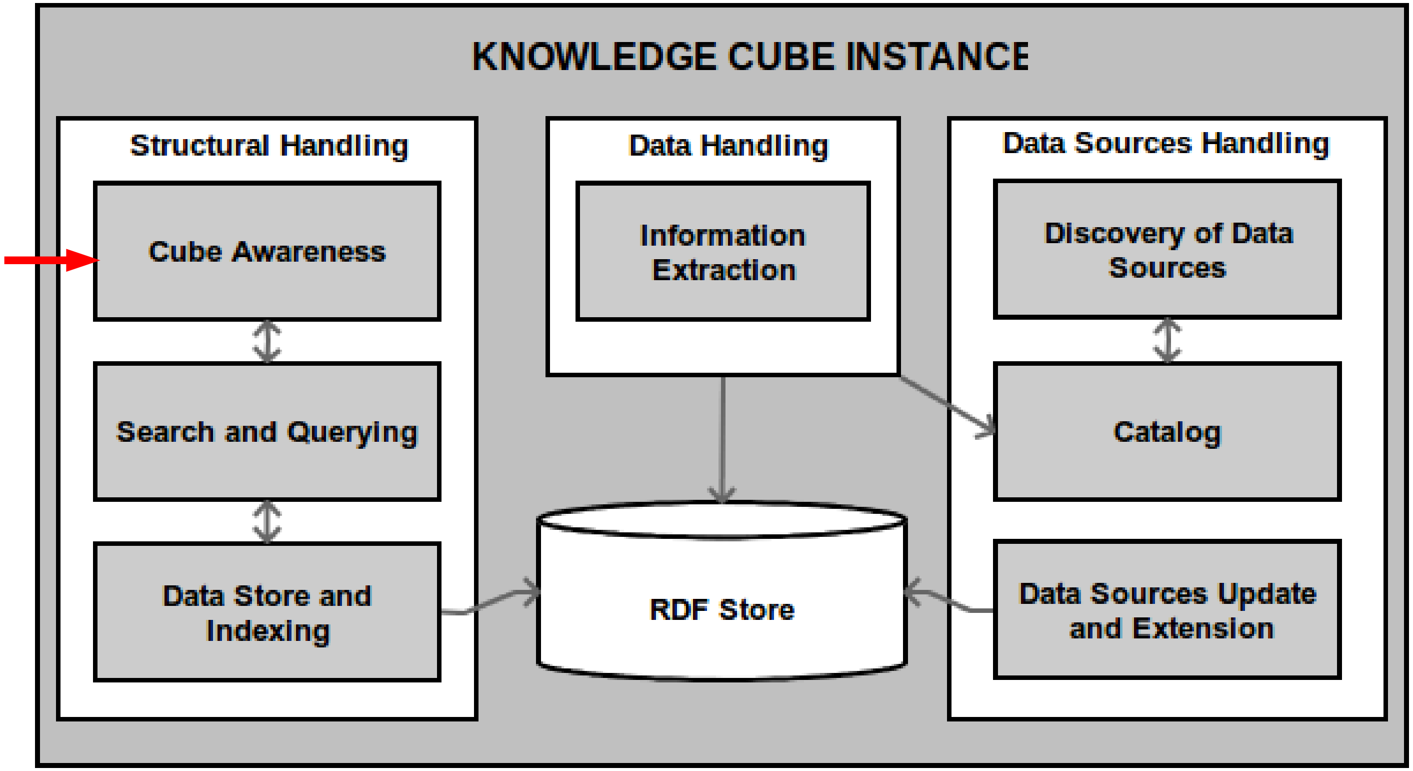
- **Data Sources Update and Extension**
  - Integrate newly acquired data in an unsupervised manner
    - **Ex:** A data source indicates that a certain <Subject> (ex: Bush) is no longer president

Architecture



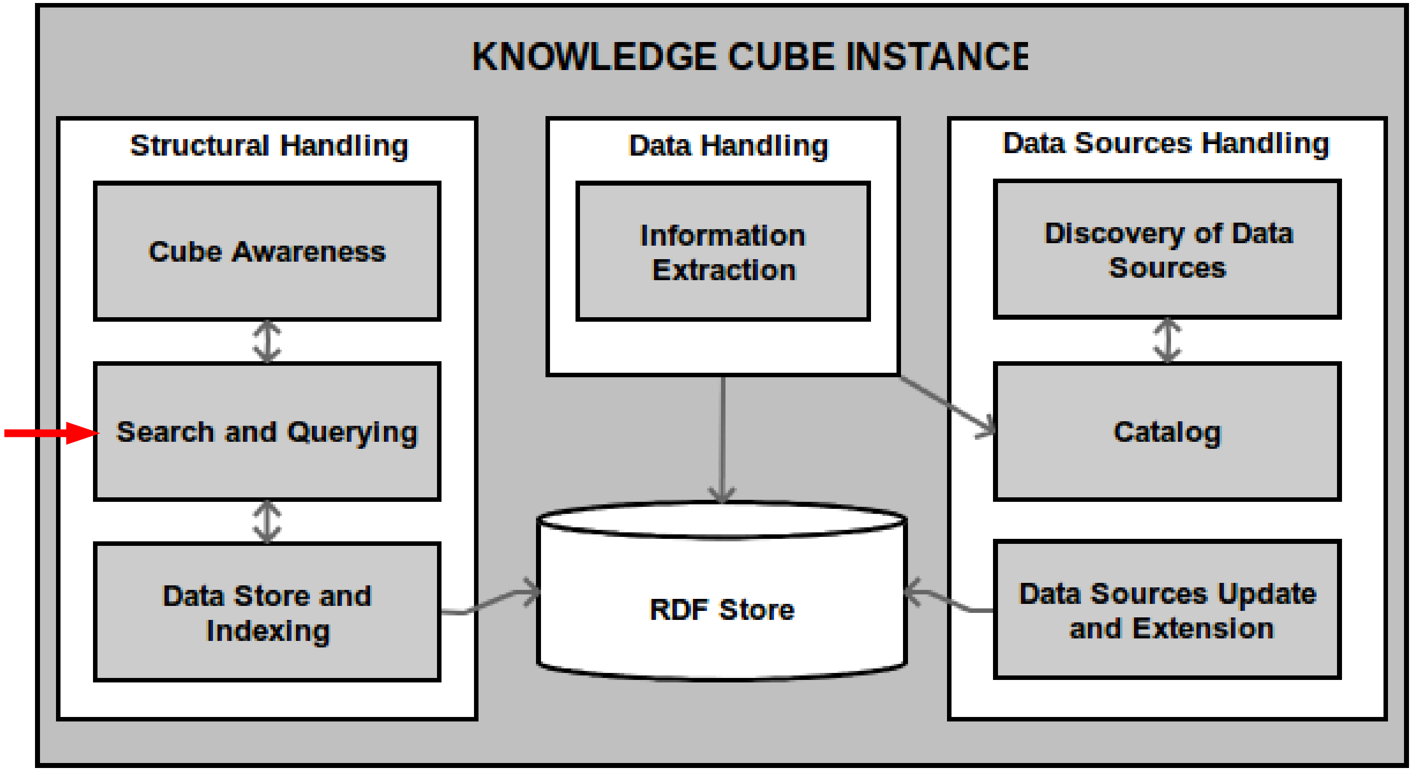
- **Information Extraction**
  - Employs text analysis to extract and learn from structured and unstructured text
    - **Ex:** Extract <Subject,Predicate,Object> using unsupervised techniques

Architecture



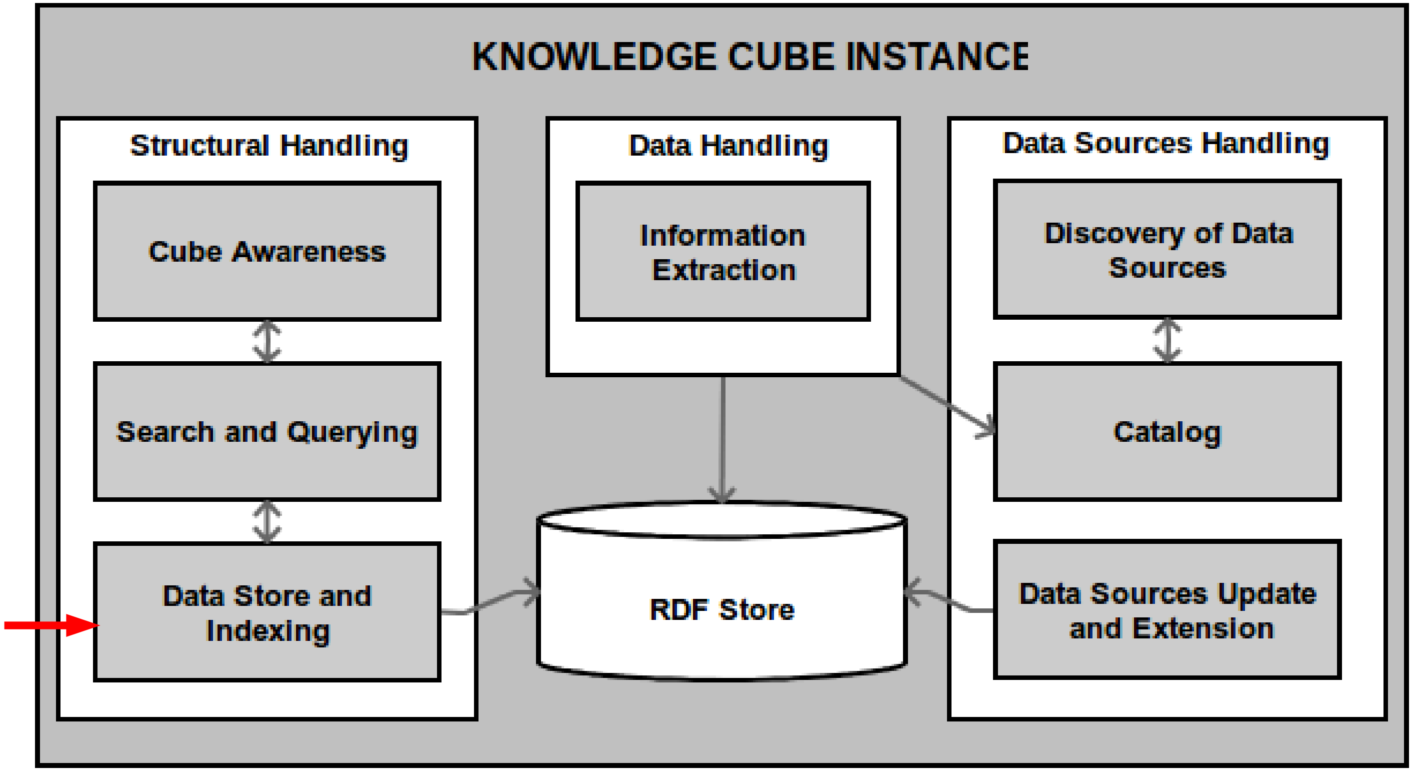
- **Cube Awareness**
  - Provides structural or data-level updates to the cube
    - ***Ex:*** New cube has been created that handles data about basketball

Architecture



- **Search and Querying**
  - Constructs to understand the semantics of the query terms
    - **Ex:** Providing a SPARQL and Geo-SPARQL query capabilities

Architecture



- **Data Store and Indexing**
  - Efficient and scalable storage and indexing mechanisms over RDF triples
    - **Ex:** Connect to Relational DBMS, Triplestore or other RDF data store

Challenges

- **Semantic Interpretation**
  - Interpretation of ambiguous or imprecise data
    - **Ex:** Weather data (Is the data in Celsius ? Fahrenheit ?)
- **Uncertainty**
  - Attach a truth value to the extracted data
    - **Ex:** 90% confident that this is Bush (President) and not Bush (Band or Plant)
- **Data Partitioning Scheme**
  - Defining an efficient scheme for partitioning
    - Based on Named Entity Type (Sport,Organization,Name)? Time? Spatial?

Challenges

- **Storage and Indexing**
  - Choice of **storage scheme**
    - **Ex:** Triplestores, Vertically-partitioned tables, schema-specific systems, Other?
- **Communication among Cubes**
  - Define a **protocol** that considers the overhead of contacting and retrieving content from other knowledge cubes
    - **Ex:** How many cubes should we contact to give a precise answer ?
- **Data Change Frequency**
  - Identify when a knowledge cube updates its Linked Data
    - Hourly ? Weekly ? On-Demand ?

Summary

- An architecture driven by **data semantics** called Knowledge Cubes
- The data semantics includes **spatial**, **temporal**, **topical** and **contextual**
- Knowledge cubes founding principles include **structural evolution**, **temporal evolution** and **analytic distribution**
- A Knowledge cube aggregates and responds to queries only **relevant** to its data semantics