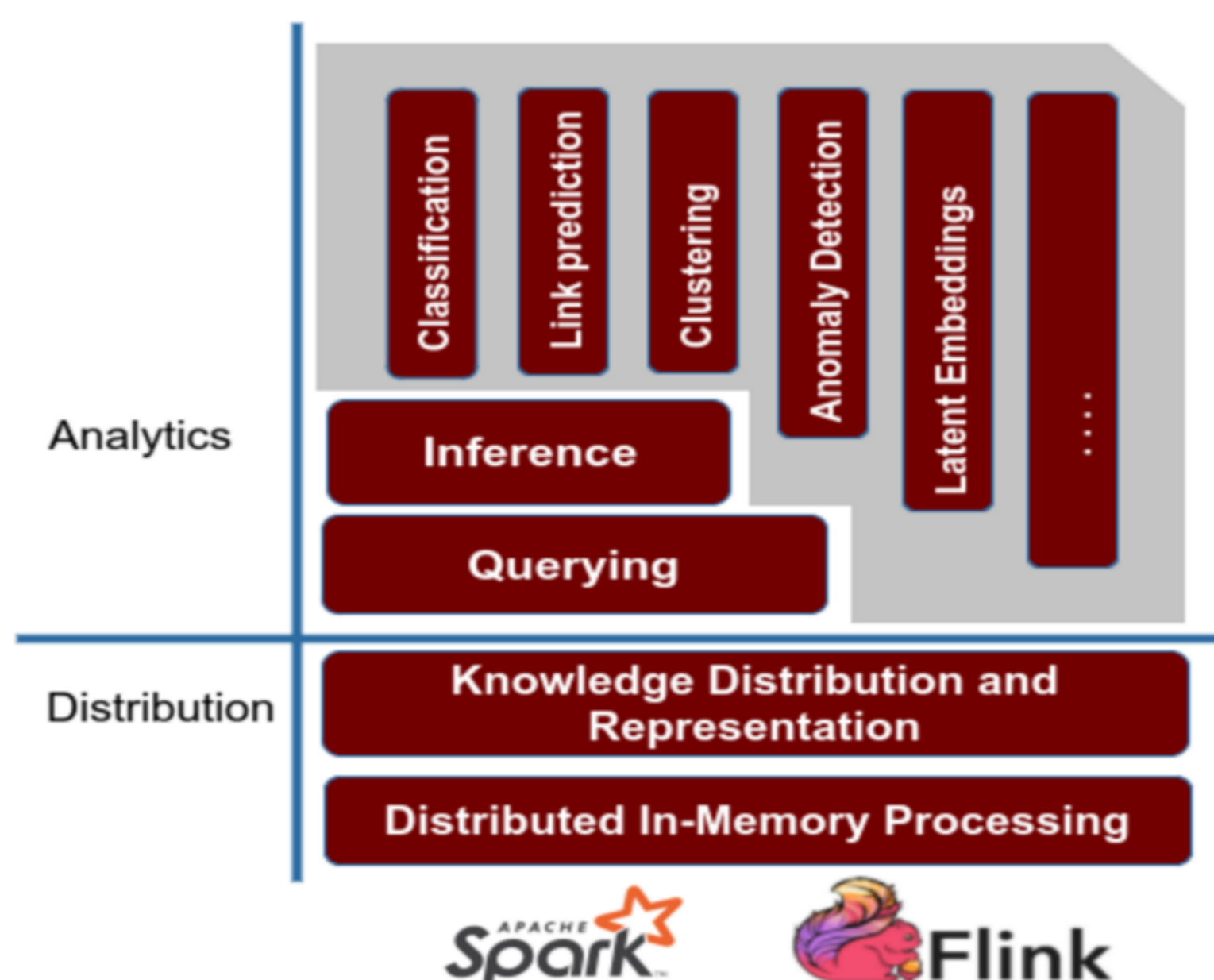


Anisa Rula*, Gezim Sejdiu**, Andrea Maurino*
*University of Milano-Bicocca, **University of Bonn

Introduction

- Creating and managing large-scale RDF datasets has been the key to success for many applications, such as semantic search, query answering and machine reading
- The quality of such knowledge bases plays a fundamental role in large-scale data consuming applications.
- Challenges
 - Traditional techniques of quality assessment for RDF datasets are not adequate to assess the quality at large-scale.
 - These approaches mostly fail to capture the new dimensions (the four Vs) of big data.

Sansa Framework



Approach

We are working on a novel approach for computing RDF dataset quality metrics and implement it using an efficient framework for large-scale, distributed and in-memory computations.

Integrate our approach into the SANSa framework which is a big data processing engine for scalable processing of large-scale RDF data.

Combining Sansa with Data Quality Dimensions

The computation of the set of quality metrics in our approach is performed as follows:

1. Defining quality metrics parameters
2. Retrieving the RDF data
3. Parsing and mapping RDF into the main dataset
4. Quality metric evaluation

