ParADS

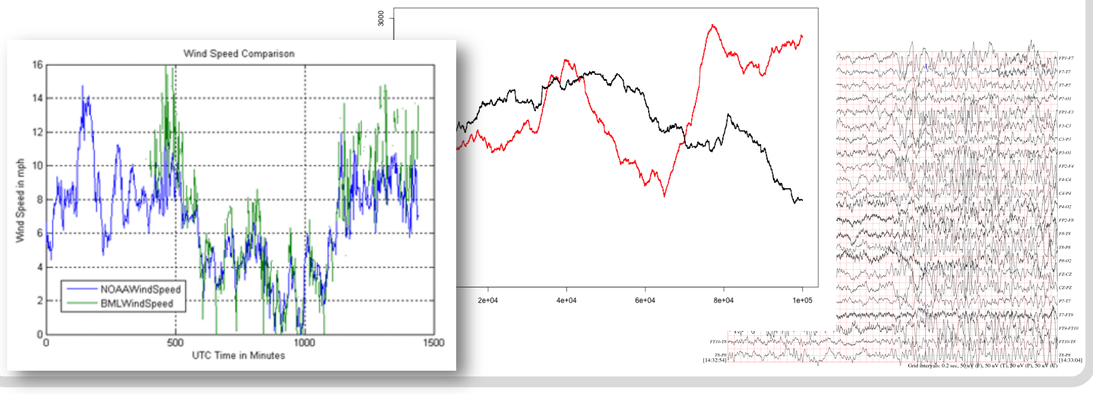
Scalable Indexing of Very Large Data Series Collections Using Modern Hardware

Botao Peng
Paris Descartes University
botao.peng@hotmail.com

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

Motivation

data in many domains are in the form of **data series**
several data series collections in the order of **multi-TBs**
across different domains: astronomy, finance, IoT, etc.



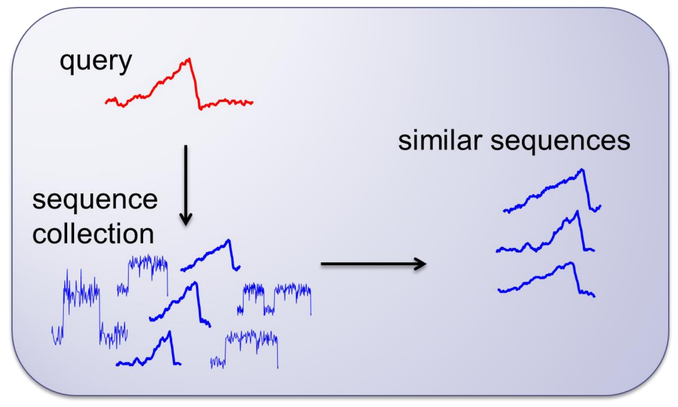
Problem

how to perform **complex analytics** on **massive collections** of sequences

- subsequence similarity search
- classification
- clustering
- frequent patterns
- outliers
- …

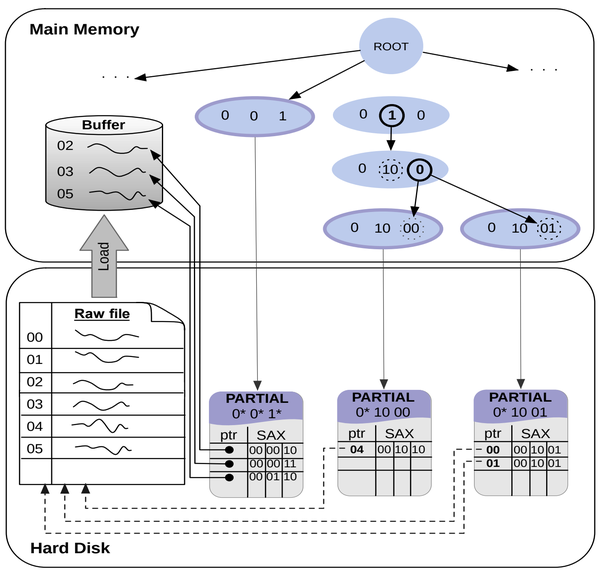
Similarity Search

a key operation of all these analytics tasks is **similarity search**

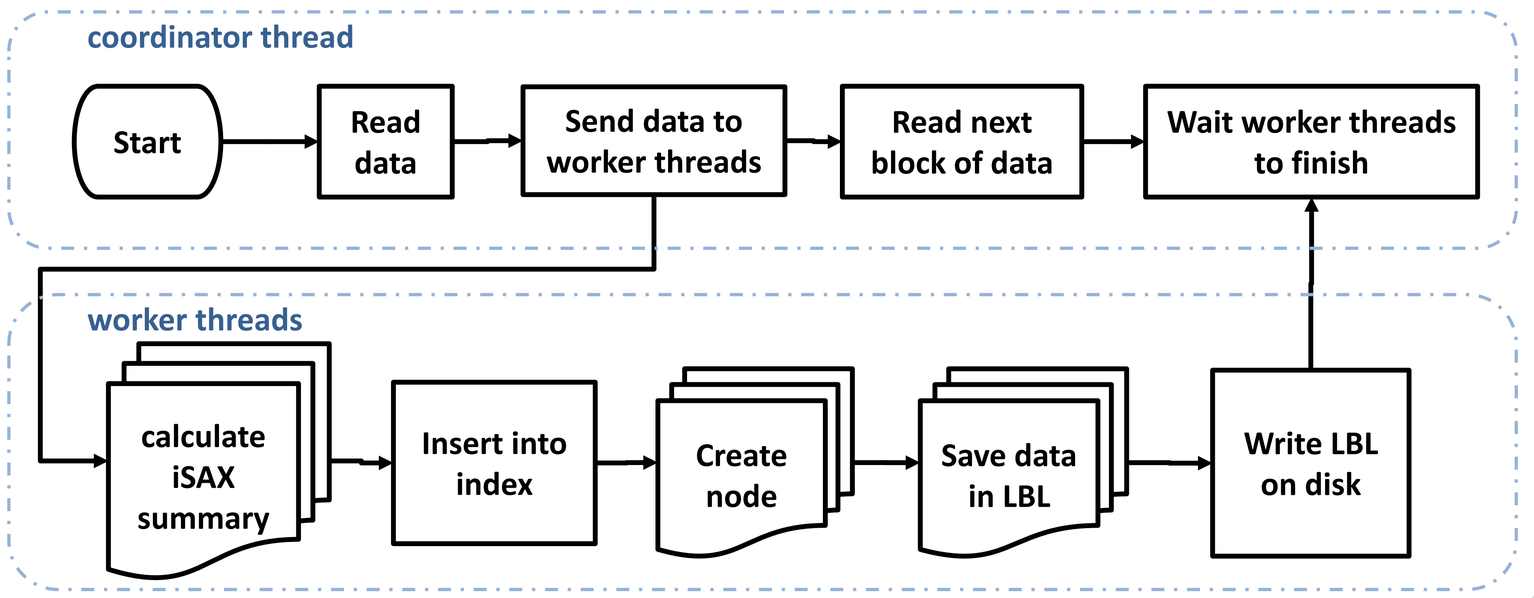


Adaptive Data Series Index (ADS+)

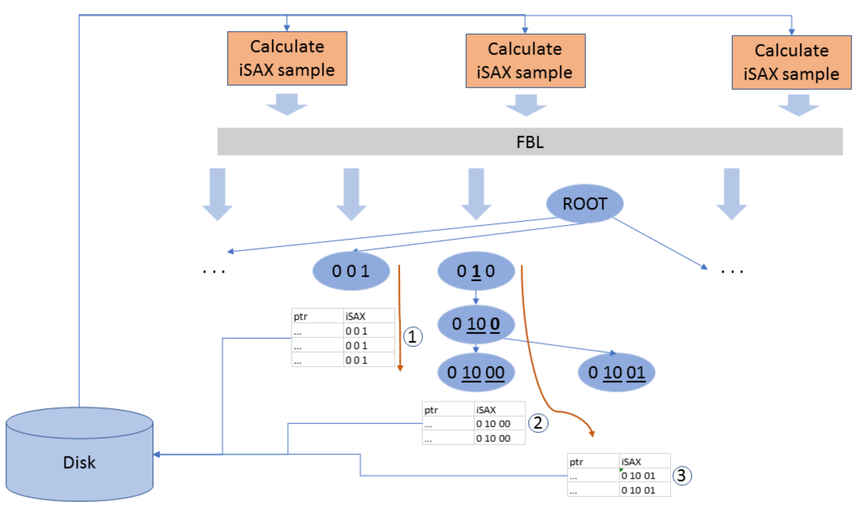
current state-of-the-art technique for fast similarity search queries



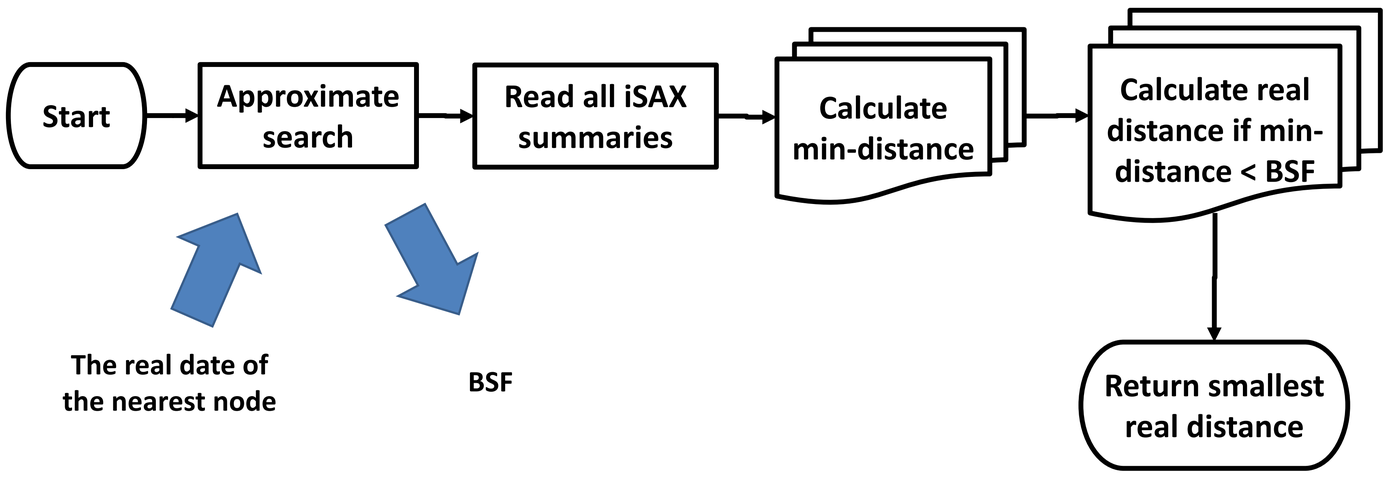
Parallel Index Creation (using multi-cores)



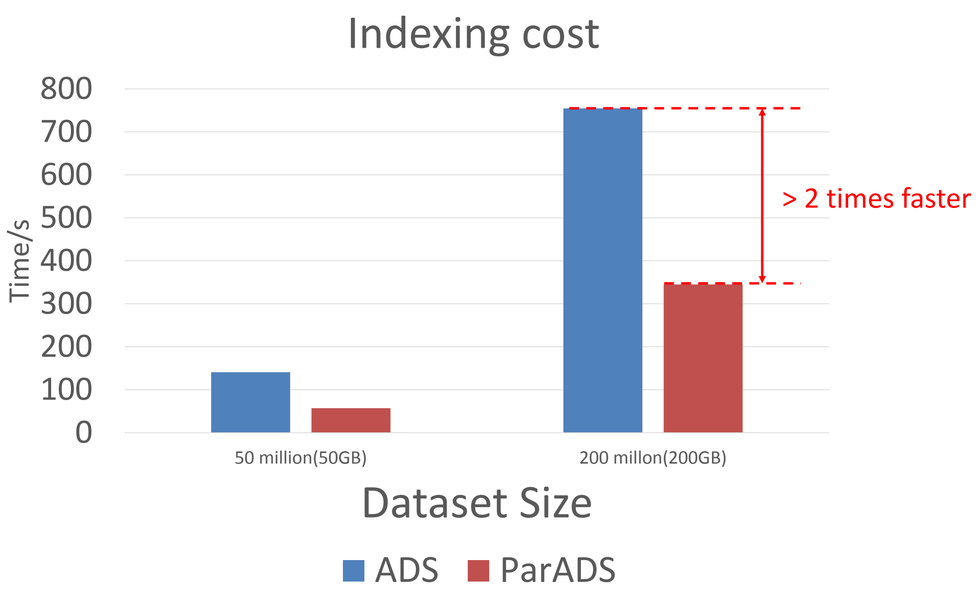
Indexing Architecture

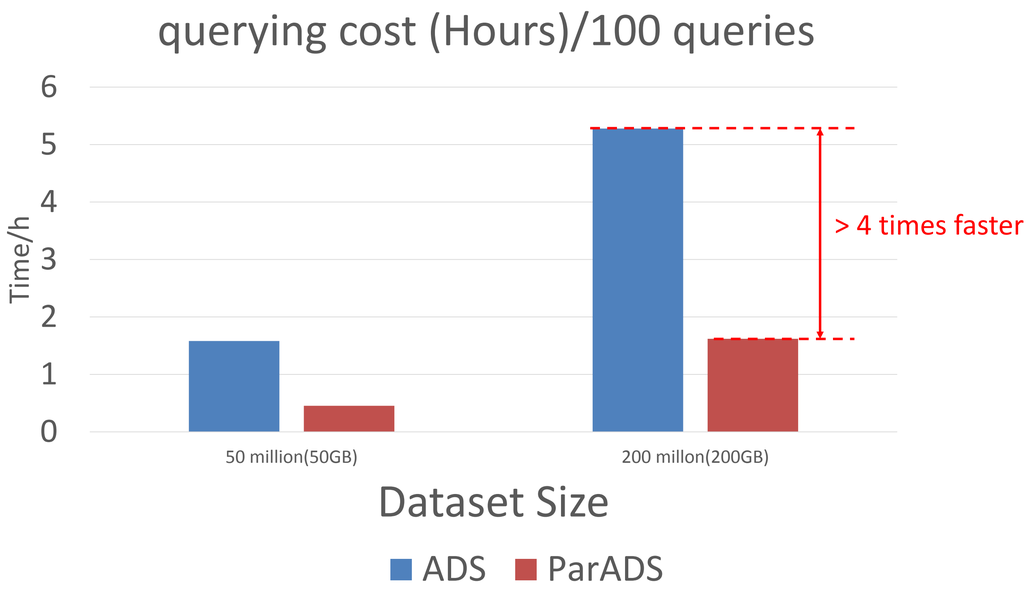


Parallel Query Answering (using multi-cores)



Experimental Results





References

ADS: The Adaptive Data Series Index. VLDBJ 2016

Big Sequence Management: A Glimpse on the Past, the Present, and the Future. LNCS, 2016

Query Workloads for Data-Series Indexes. KDD 2015

RINSE: Interactive Data Series Exploration. VLDB 2015

UNIVERSITÉ PARIS DESCARTES — MEMBRE DE USPC Université Sorbonne Paris Cité

LIPADE Laboratoire d'Informatique PAris DEscartes

google: diNo Paris5
http://dino.mi.parisdescartes.fr/