Horme

Random Access Big Data Analytics

Guangchen Ruan, Beth Plale; *Milinda Pathirage (Presenter)*

School of Informatics and Computing
INDIANA UNIVERSITY

Introduction

Hathitrust Research Center (HTRC)

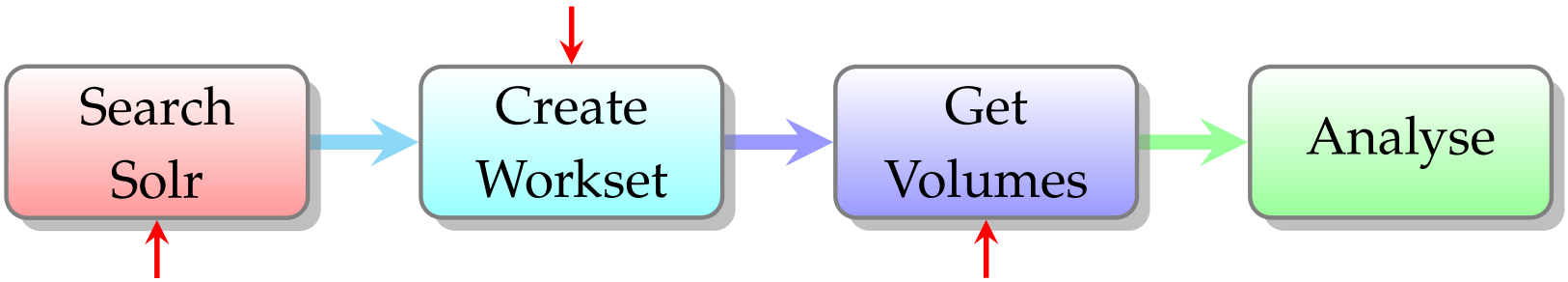
Research arm of **Hathitrust Digital Library** that develops and maintain infrastructure for enabling computational access to **14 million** digitized volumes.

Mission

Enable researchers world-wide to accomplish tera-scale text data-mining and analysis

HTRC Analytics Workflow

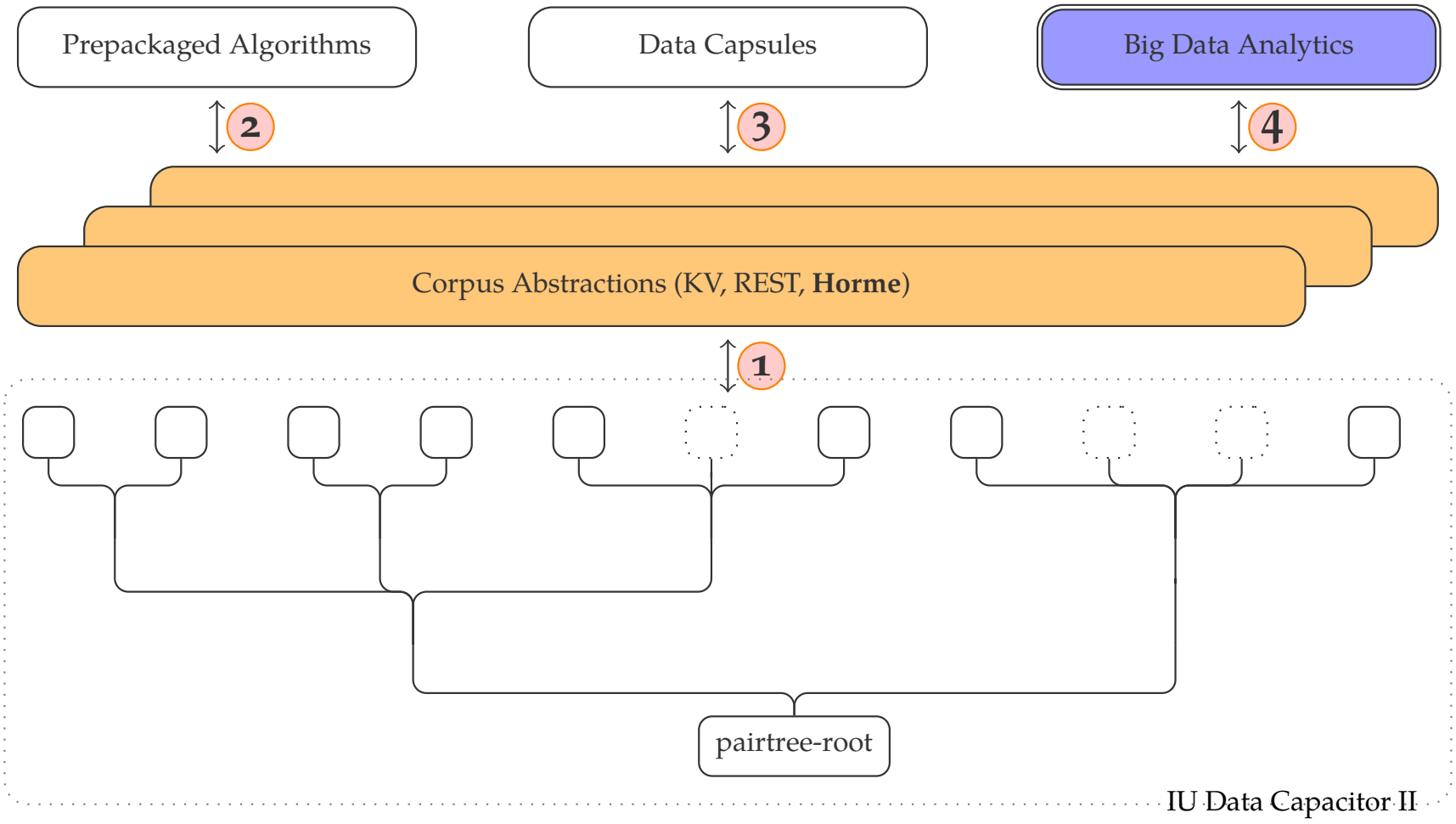
Workset is given as an input to analytic sub-systems that downloads volumes from storage sub-systems integrated to them



Select a subset of data for analysis

Random access to the corpus

HTRC Analytics Infrastructure



Motivation

Hadoop based Text Data Mining on HPC

- Data is stored in a *parallel file system (PFS)* such as Lustre connected to compute nodes via network
- Often data get staged to the scratch space in compute nodes (e.g. HDFS data nodes) from PFS before the actual computation
- Results get copied back to PFS after job completion
- **High data staging overhead**
- HDFS on HPC is limited by scratch space capacity of compute nodes

Random Access Text Data Mining on HPC

- Use Solr search or some other means to narrow down the set of digitized texts to analyse
- Apply text mining techniques such as topic modeling on the selected subset
- Often this subset is randomly distributed across the corpus (Around 14 million volumes in HTRC)
- HDFS performs poorly in random access use cases
- HBase is good for random access, but needs to deploy external to HPC compute nodes due to transient nature of HPC jobs
- HBase over PFS is not optimal

Horme