DATA MANAGEMENT ISSUES IN BIG DATA ANALYTICS

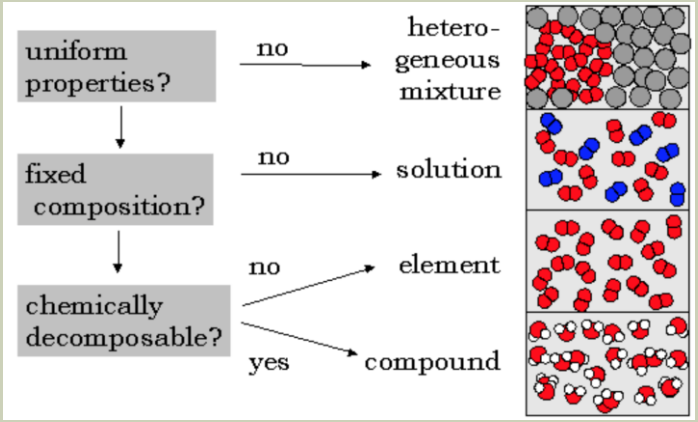
王晓阳 X. Sean Wang
School of Computer Science
Fudan University
xywangCS@fudan.edu.cn

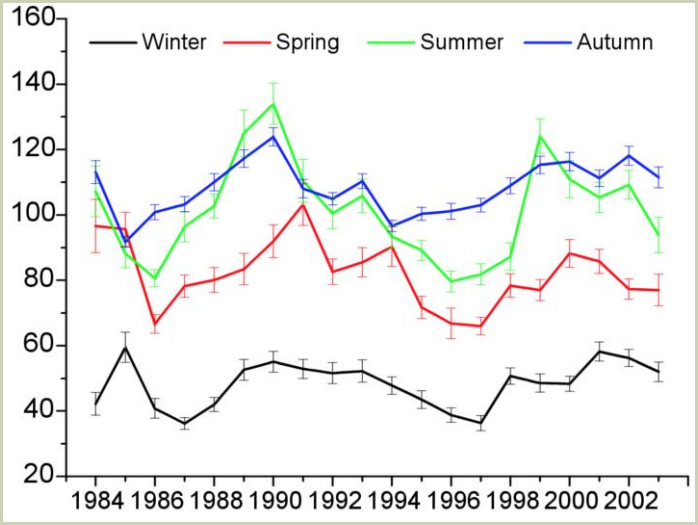
OUTLINE

- Defining “Big Data Analytics”
- Two technical issues
  - Semantic integration
  - Data placement on cluster of compute nodes
- Conclusion

BIG DATA ANALYTICS

- It's all about **prediction models**
  - Classification models
  - Clustering
  - Time series prediction models
- Or even **simple statistics**
  - Summary statistics
  - Projection and plotting
  - … for human analysts to "spot" prediction models





BIG DATA VS "SMALL" DATA WHAT'S THE DIFFERENCE ?

- Volume, Variety, Velocity
- Variety issue
  - Semantic variety problem
- Volume issue: Divide and Conquer
  - Data placement problem

SEMANTIC VARIETY

- Prediction problem:

  Given: $f(reality_1)$, …, $f(reality_{n-1})$

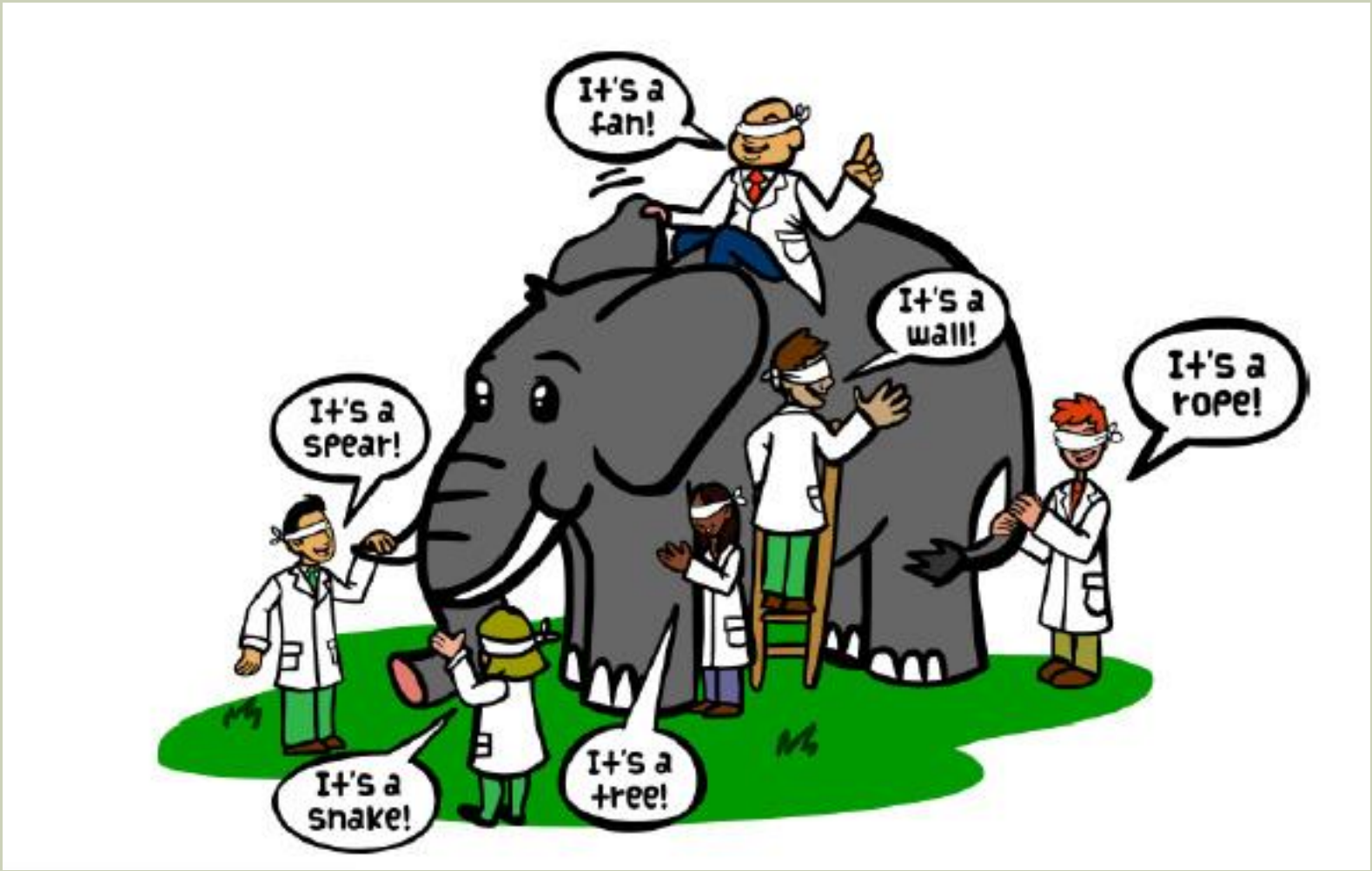
  Question: $f(reality_n)$=?

SEMANTIC VARIETY

- In practice, we often have:
  - $f_1(reality_1)$, …, $f_1(reality_{n-1})$
  - …
  - $f_k(reality_1)$, …, $f_k(reality_{n-1})$
- We need to predict
  - $f(reality_n)$=?
  - Through an understanding of $reality_n$

- Each *f* is a "projection" of the "reality"
  - Probabilistic/noisy "projections"
- And we often don't really know the exact relationships among the *f*'s

THE BLINDS & THE ELEPHANT



SEMANTIC VARIETY CHALLENGES

- Traditionally:
  - Data integration/fusion
  - The hope is to derive *f(reality)* from f_i(*reality*)
  - Then to use f(*reality*) to understand *reality*
- Generative model
  - Assumptions of how observations are generated
  - Discover the *reality* from the observations
    - Each observation is used to augment the view of the *reality*
  - *Sort of the "optimization" problem*

LESSON FROM INFORMATION RETRIEVAL

- Important: Generative model
- Words: *w*, Documents: *d*, Topics: *c*
- We know a sample of $P(w,d)$
- We know a sample of $P(w|c)$
- We want to predict $P(c|d)$

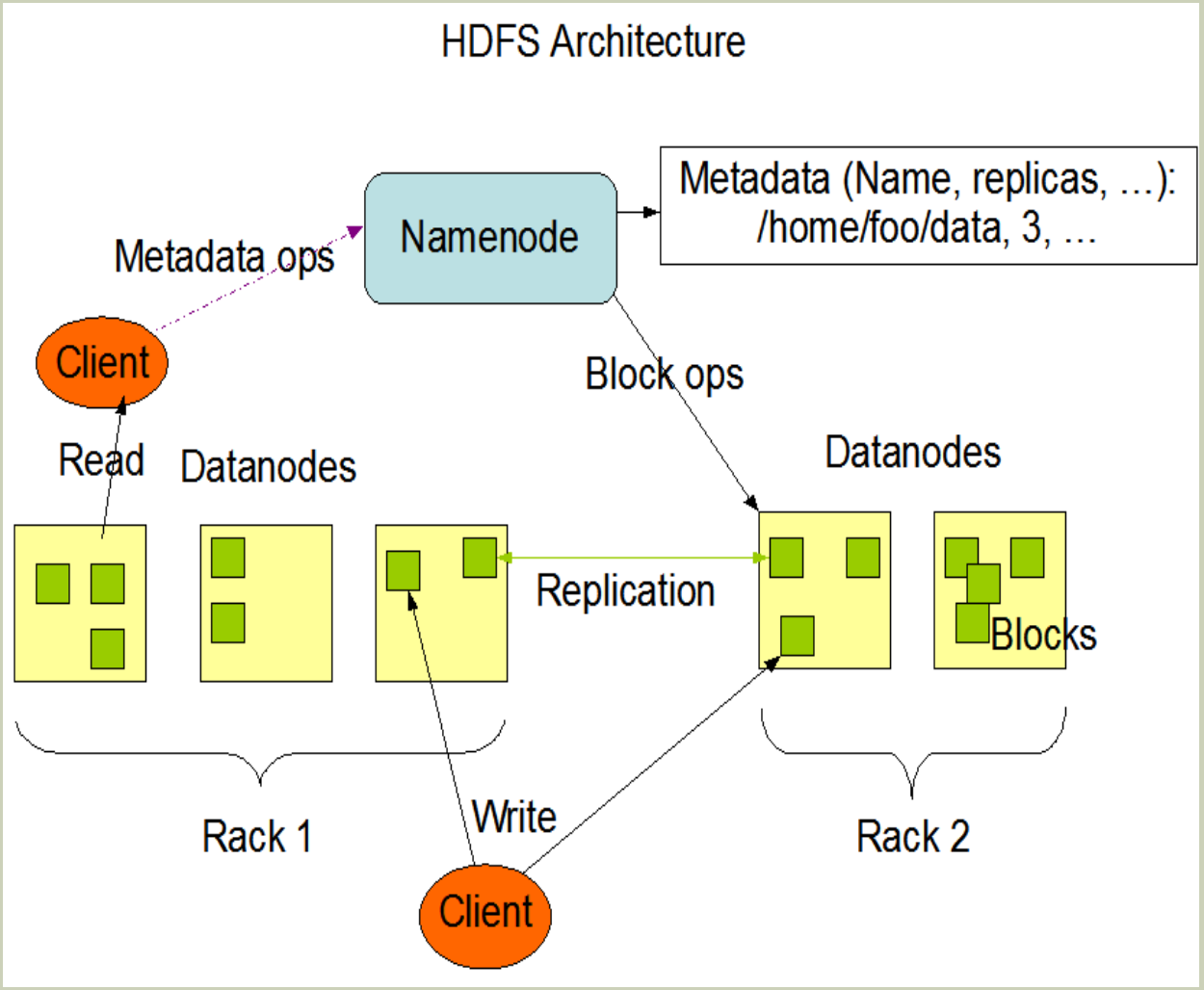
$$P(w,d)=P(d) * \sum_c (P(c|d)* P(w|c))$$

SEMANTIC VARIETY CHALLENGES

- Data markets challenge
  - How do I know which data sets to acquire in order to do my job?
  - Time/money challenge
- Data set purchase based on *description or sample?*

DATA VOLUME CHALLENGE

- Data scan
- Compression
- Load balancing



BIG DATA ANALYTICS

- New requirement: *Selective data scan*
  - Select a part of the data based on the requirements
  - Perform analysis on the subset of the data

PROBLEM & SOLUTION?

- HDFS is best for data scan
- Data scan is not “selective”
- Indexing often doesn’t help

- Solution: Data partition

- Worry about: load balancing

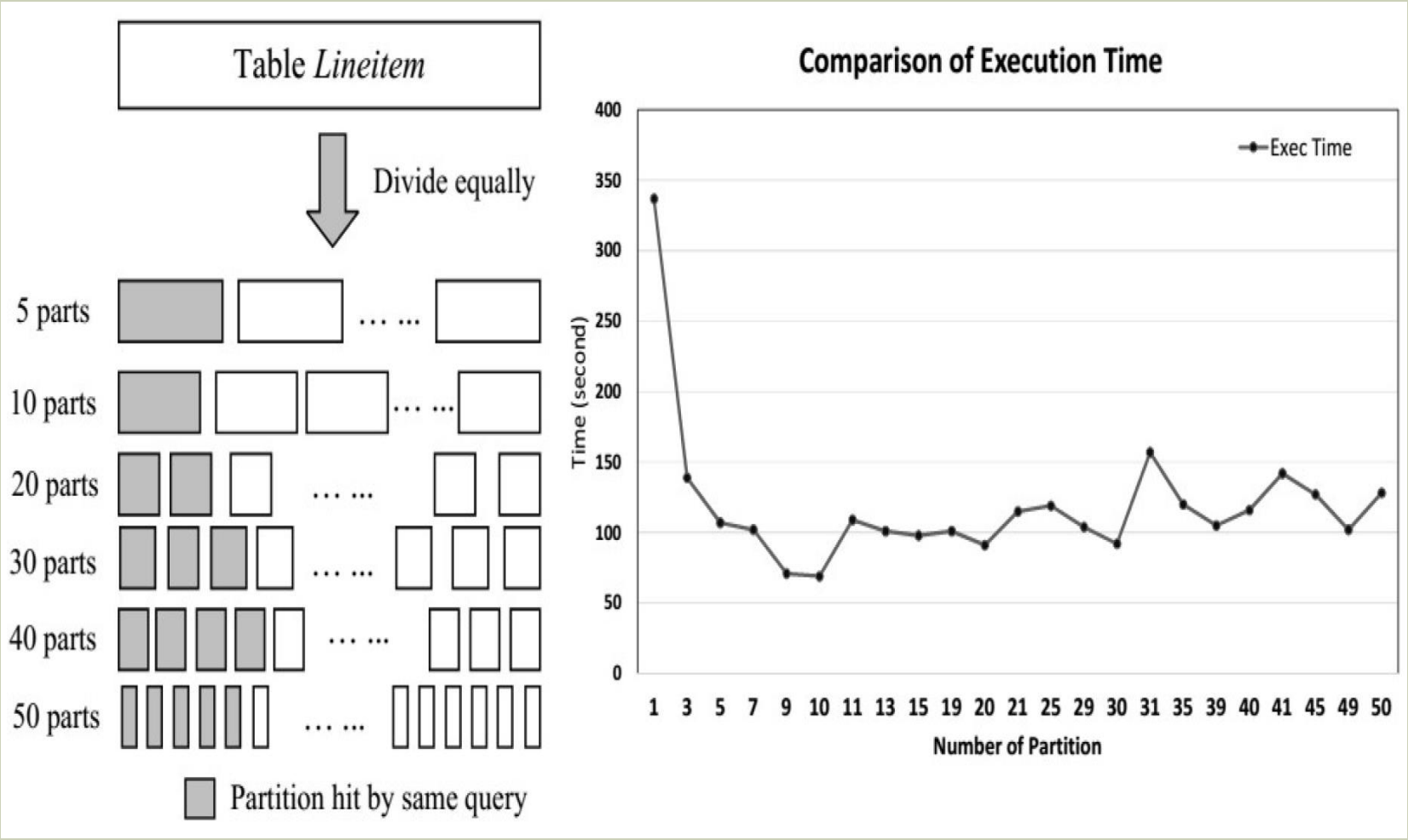
RESEARCH QUESTIONS

- With known workload, what kind of file structure can achieve the lowest time cost?
- Define cost model to predict the cost incurred by a query. Decide the file structure with the lowest cost.
- How to implement co-location in HDFS?
- How to group related columns?

EXPERIMENTS & COST MODEL

- Environment:
  - Cluster: 1 master node, and 5 slave nodes
  - Dataset: TPC-H
  - Hadoop Version: 1.2
  - Framework: Hive 0.8
  - File Structure: RCFile

TO HOW "FINE" TO PARTITION?



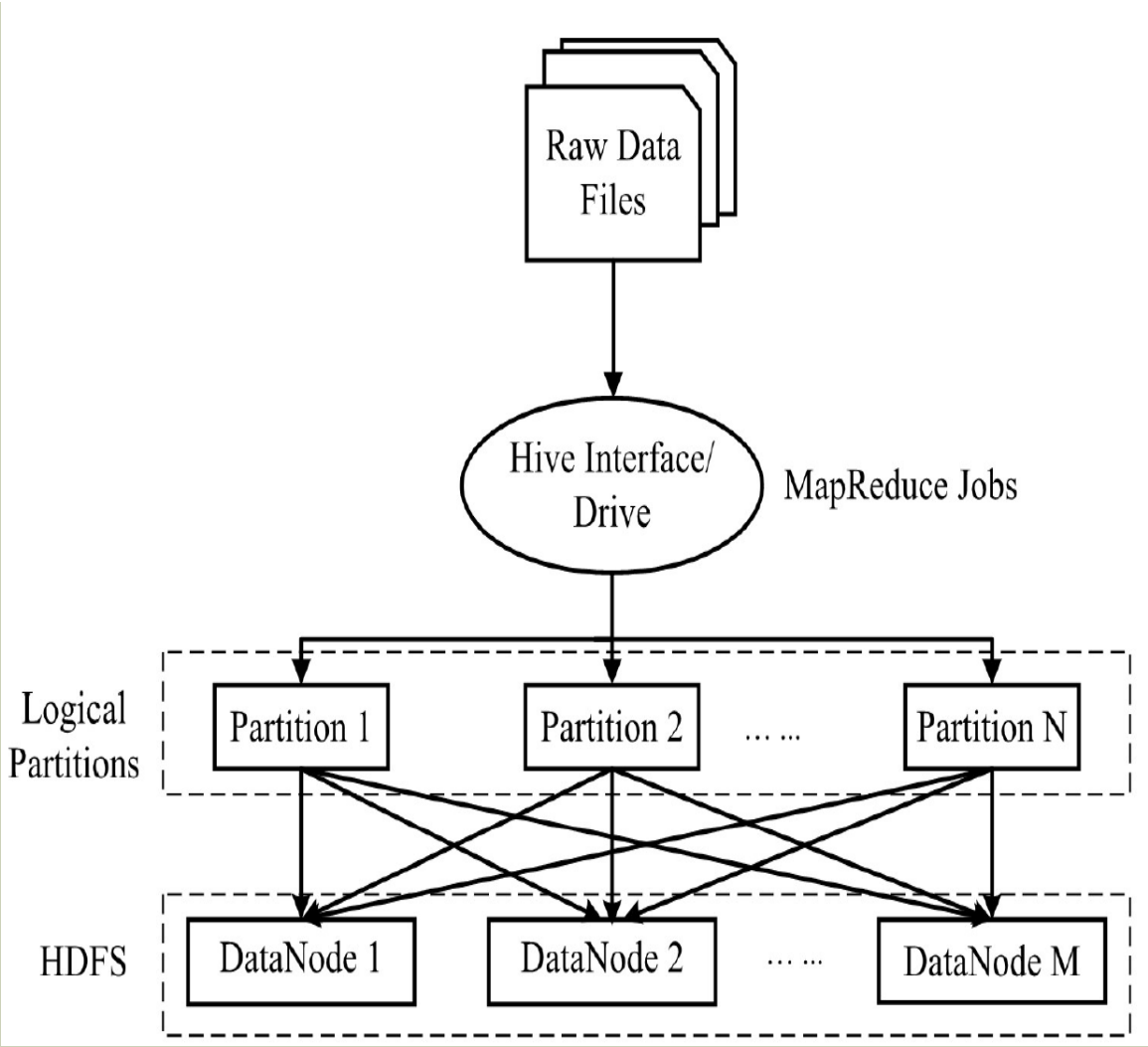
“DATABASE SOLUTION”

- Workload
  - Logical partitions
  - Based on the semantics of the data
- Cost model
- Optimization

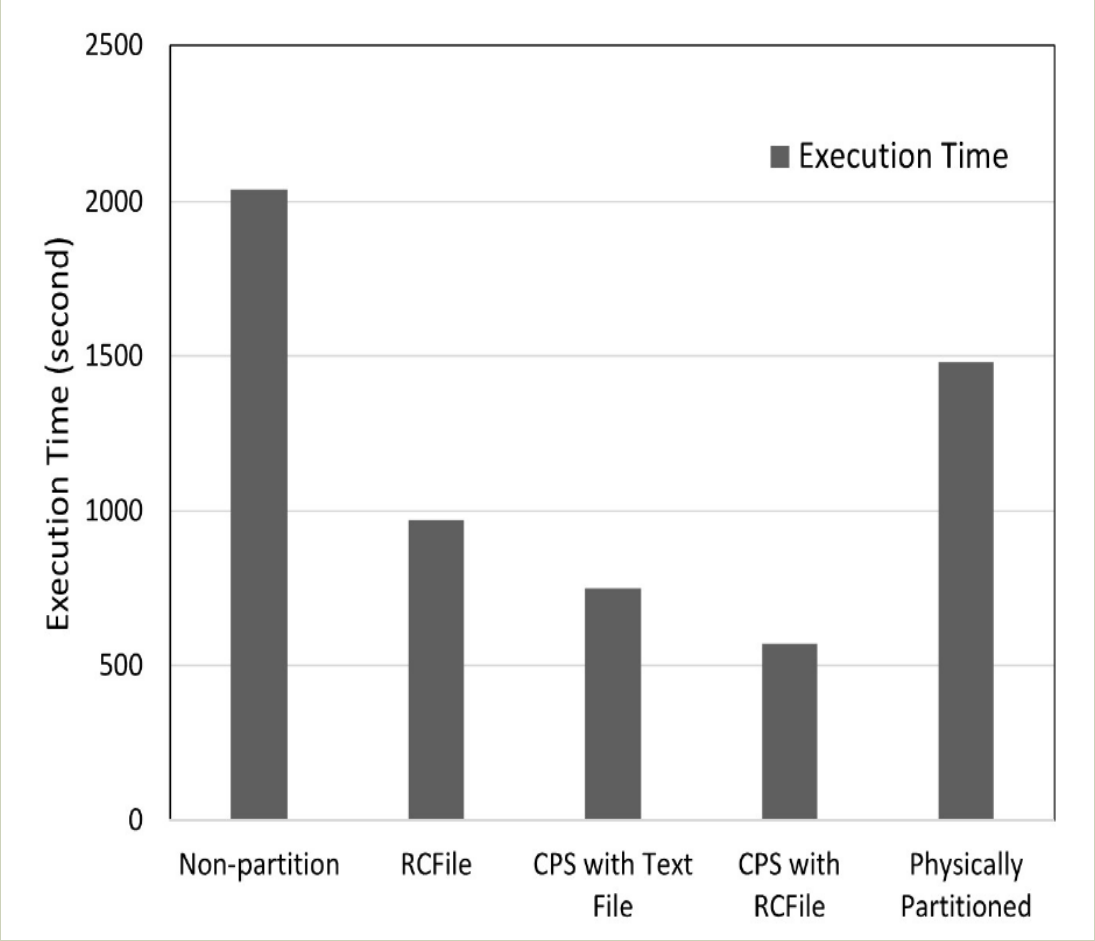
FOR SELECTION-PROJECTION-AGGREGATION QUERIES

- Partitioning based on an analysis of selection conditions
  - Finest partition: each “block” always used whole
  - Combine blocks
- Cost model
  - Data scan costs
  - Job (map/reduce) overhead costs

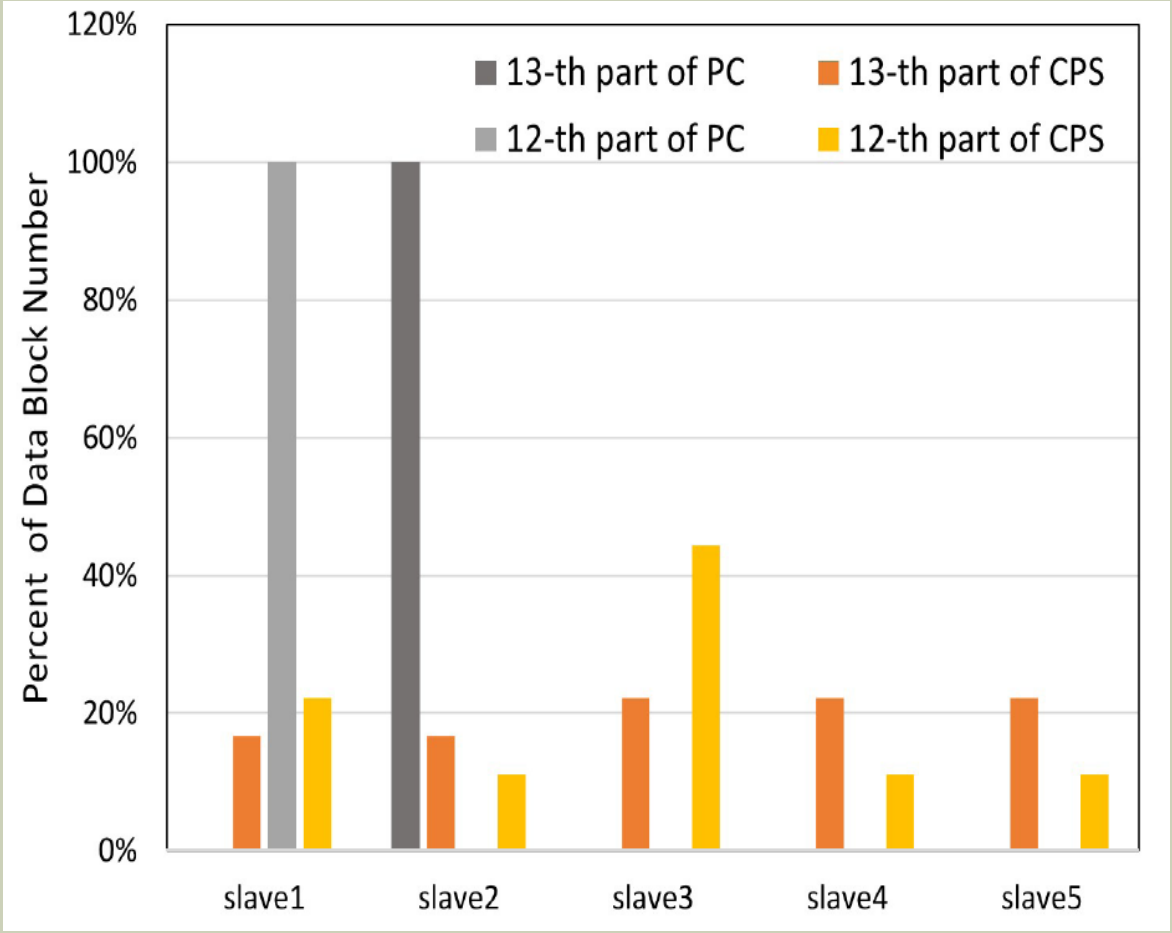
LOGICAL PARTITIONING



EXPERIMENTAL RESULTS (OUR METHOD: CPS)



EXPERIMENTAL RESULTS (OUR METHOD: CPS)



CONCLUSION

- Data semantics is central to big data analytics
  - From application level
  - to efficiency considerations

THANKS!

王晓阳
xywangcs@

fudan.edu.cn