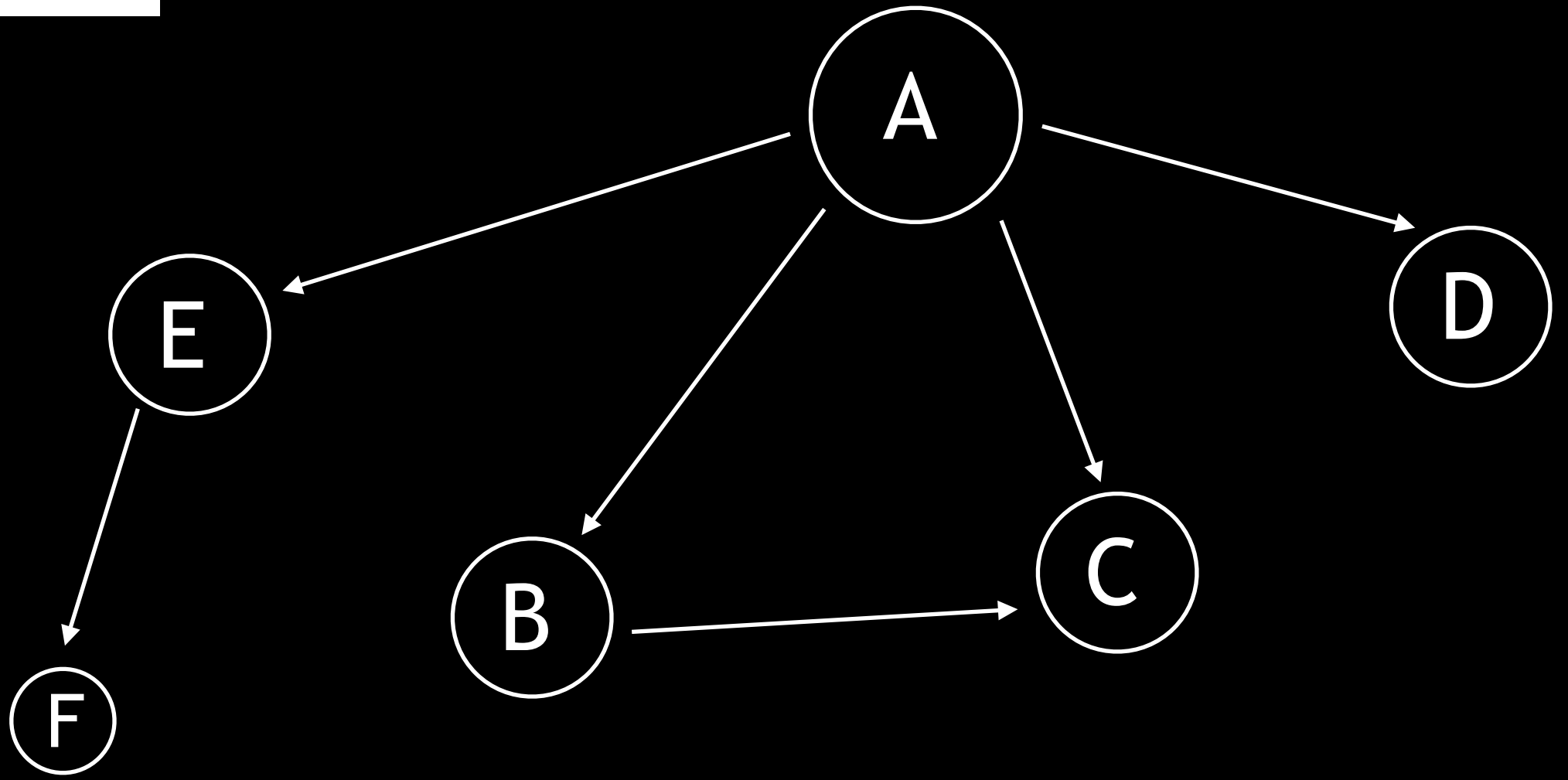
Paper Citation Analysis

Large-Scale Parallel Data Processing Project

HSIANGWEI CHAO JIEYU SHENG

1. GOAL



1.GOAL

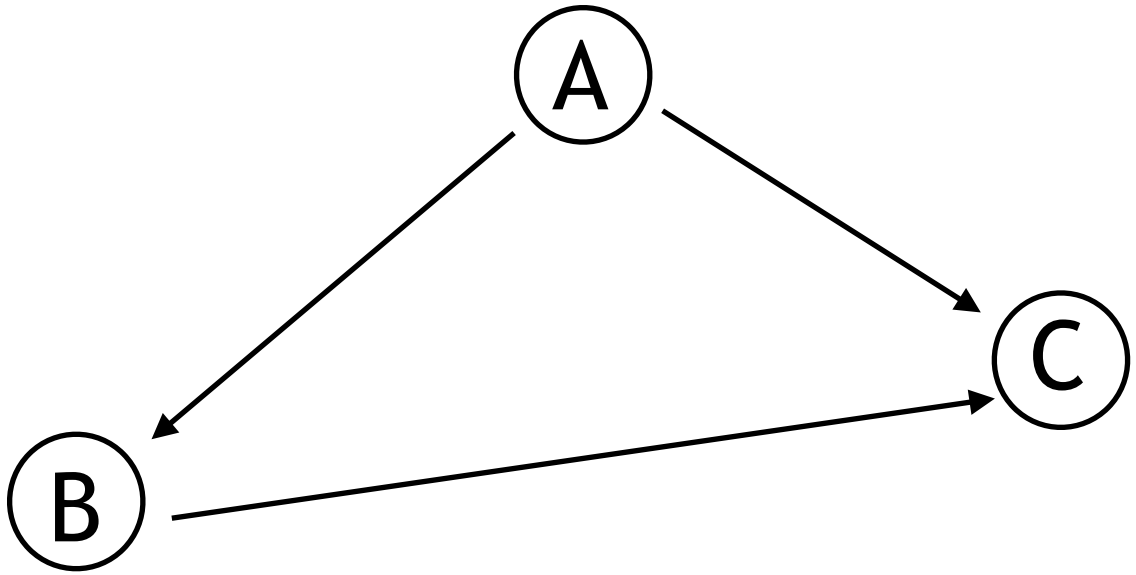
Data set	#paper	#Citation Relationship	Size
DBLP-Citation-network V5	1,572,277	2,084,019	845.4 MB
DBLP-Citation-network V8	3,272,991	8,466,859	817.6 MB
DBLP-Citation-network V10	3,079,007	25,166,994	4.39 G

2 OVERVIEW

1. Setup HBase and AWS Configuration
2. HBase as Indexing to implement Equi-Join
3. Optimize HBase Join
4. Implement ReduceSideJoin in Spark
5. Analysis the performance of each dataset and cluster

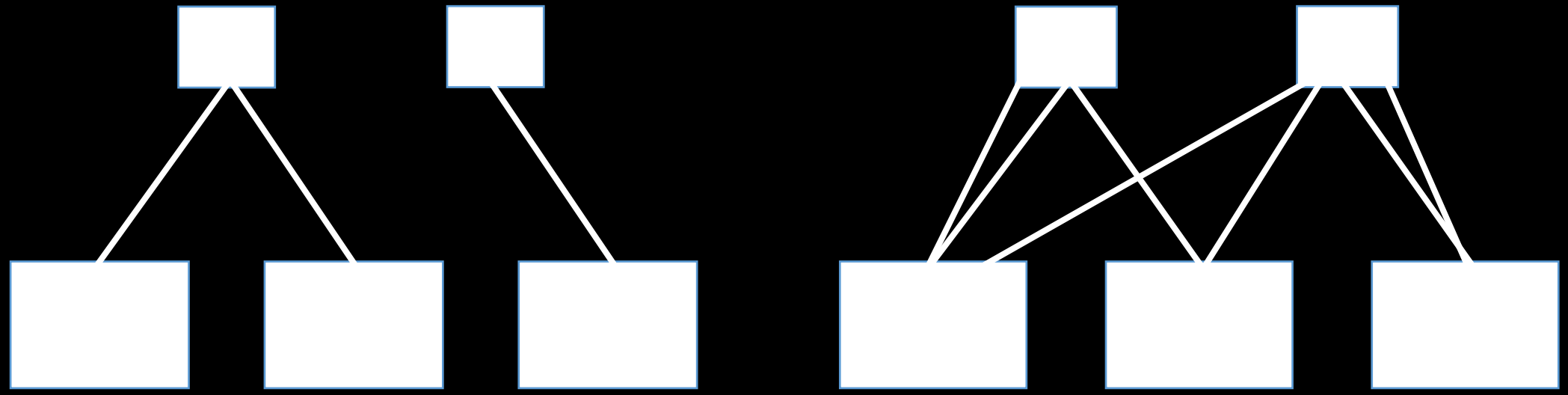
3.IMPLEMENTATION

- Set up for ReduceSideJoin: Store HDFS, Plain Json File on S3
- Set up for HBaseJoin: HBase Table with storage on S3
- Two Joins



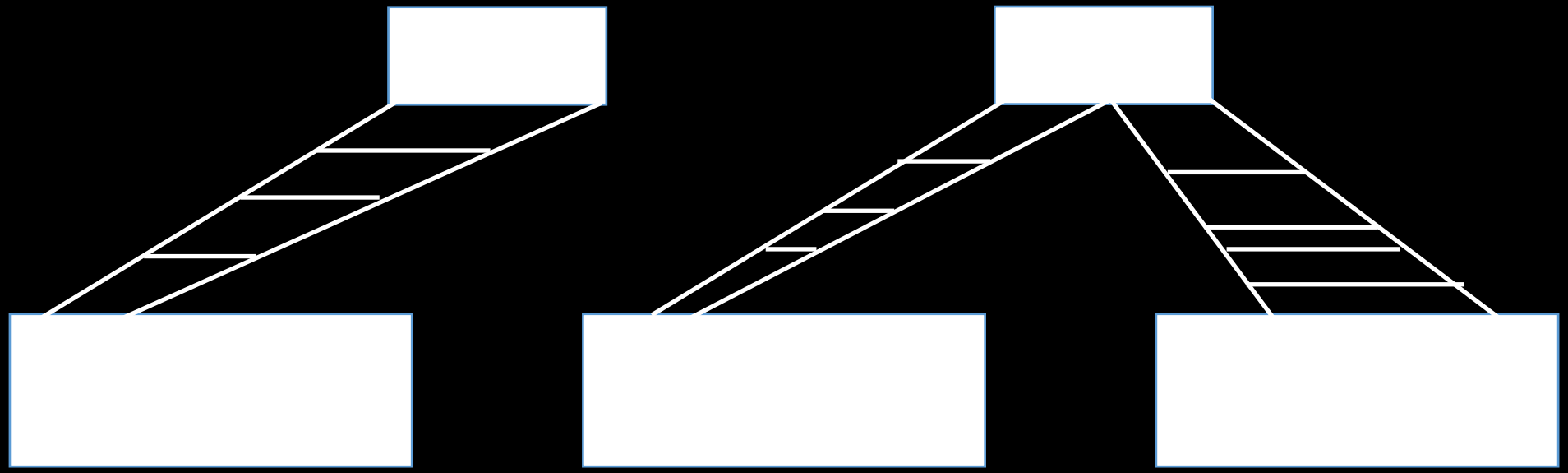
4. OPTIMIZATION

- **Use multiple gets**



4. OPTIMIZATION

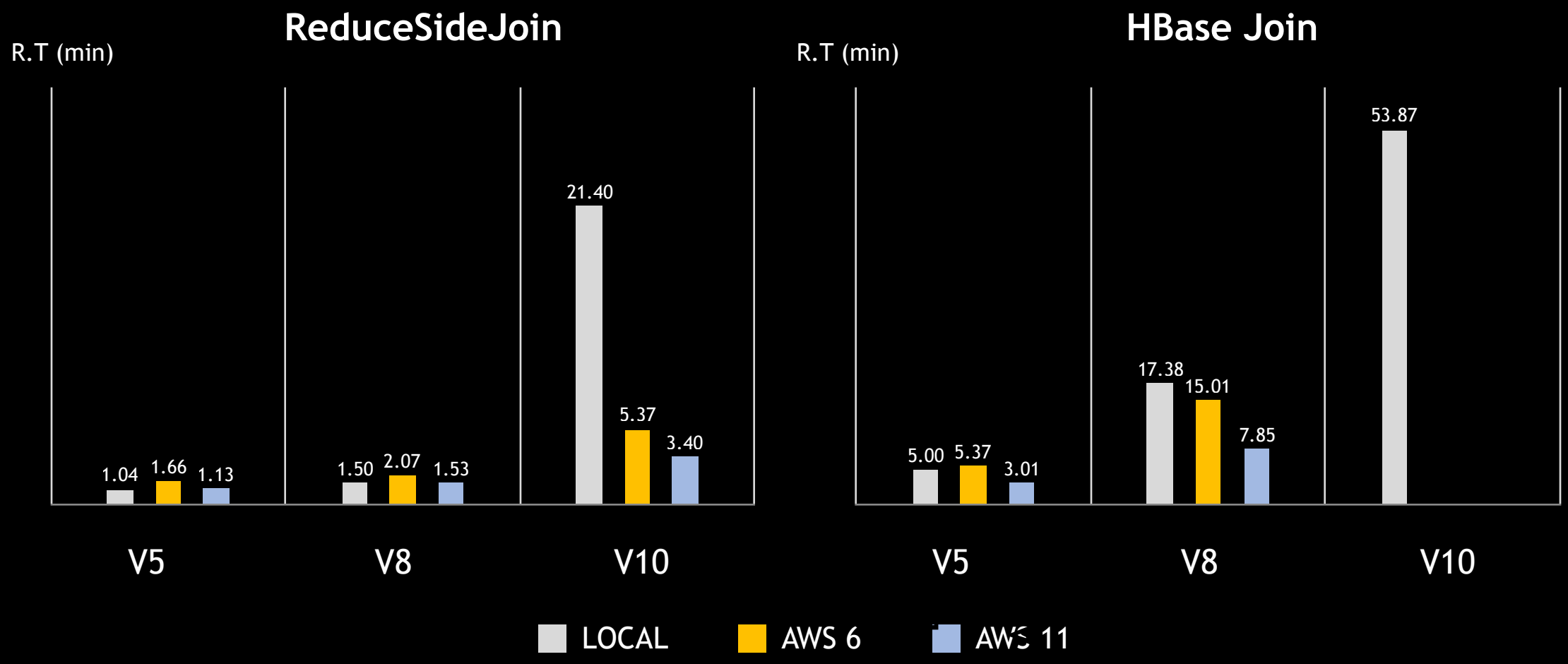
- **Use Scan**



4. RESULTS

Dataset	Result	Count	Probablity
DBLP-Citation-network V5	trangle	932315	40.05%
	total	2,327,450	
DBLP-Citation-network V8	trangle	3,405,499	39.40%
	total	8,650,089	
DBLP-Citation-network V10	trangle	12,591,863	50.03%
	total	25,166,994	

4. RESULTS


ReduceSideJoin
R.T (min)
1.04
1.66
1.13
1.50
2.07
1.53
21.40
5.37
3.40
V5
V8
V10
HBase Join
R.T (min)
5.00
5.37
3.01
17.38
15.01
7.85
53.87
V5
V8
V10
LOCAL
AWS 6
AWS 11

4. RESULTS

Speedup

	ReduceSideJoin	**HBaseJoin (scan)**	**HBaseJoin (gets)**
DBLP-Citation-network V5	1	1.07	2.25
DBLP-Citation-network V8	1.35	1.26	2.6
DBLP-Citation-network V10	1.57	-	-

5. CONCLUSION

- HBase is not suitable for Bigdataset Join
- Optimization Process
- Paper recommend system

THANK YOU