User Click Modeling based on Search Engine Log Data

Nikhil Bendre

Justin Thomas

Outline

- Source
- Background
- Dataset
- Methods
- Challenges
- Results
- Performance Benchmarks
- Conclusion

Source

- KDD Cup 2012 sponsored by Tencent Inc.
- Largest real datasets ever released publicly for competitions
- User click modeling in advertising

Background

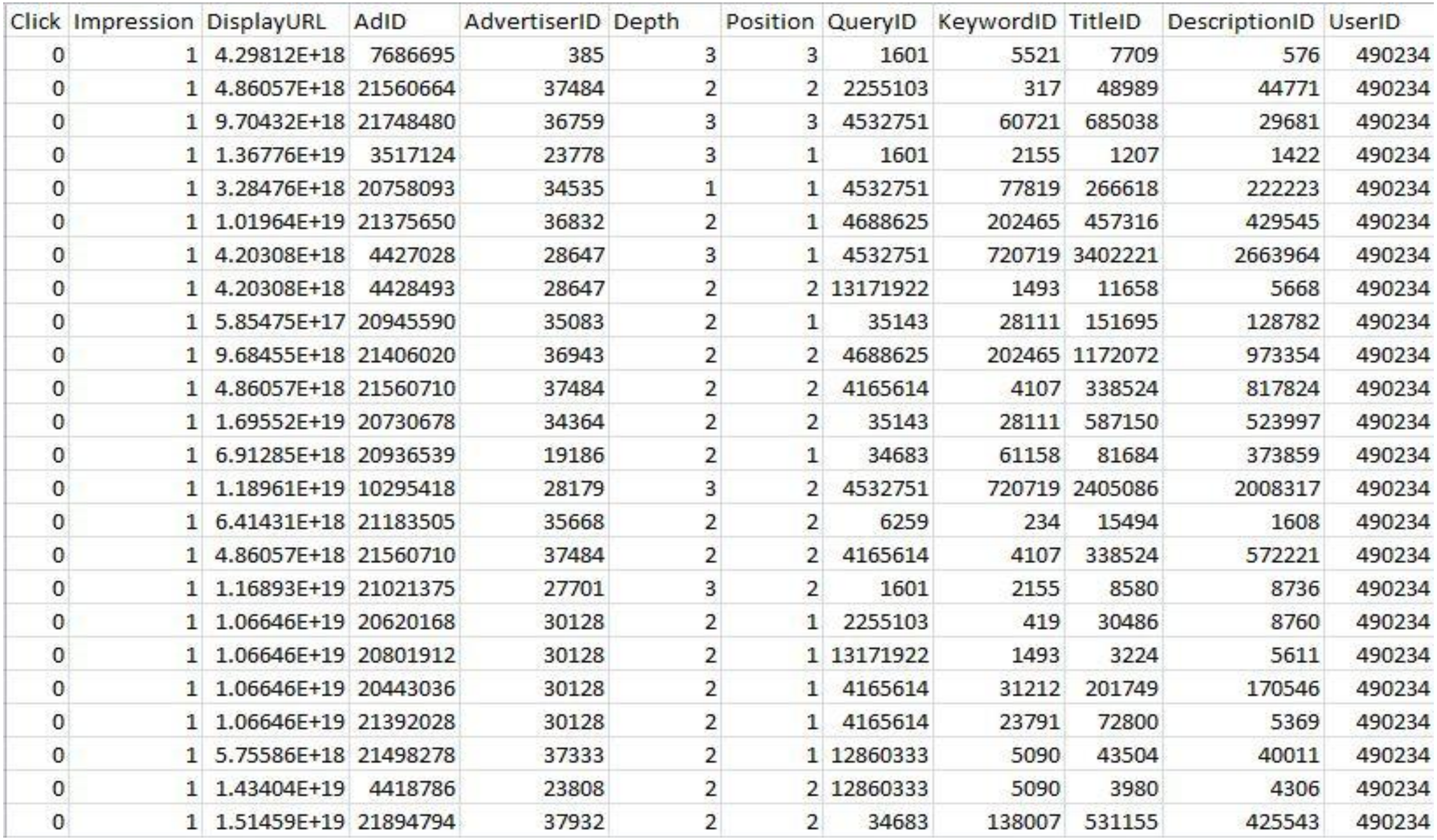
- Search Advertising
- Economic model behind it
  - Rank Ads
  - Price Clicks
- pCTR
- Session logs from Soso.com

Dataset

- Multiple data files derived from a search session
- Primary Data File
  - Training (9.87GB)
- Additional Data files
  - UserID Profile

- Click
- Impression
- DisplayURL
- AdID
- AdvertiserID
- Depth

- Position
- QueryID
- KeywordID
- TitleID
- DescriptionID
- UserID



Click	Impression	DisplayURL	AdID	AdvertiserID	Depth	Position	QueryID	KeywordID	TitleID	DescriptionID	UserID
0	1	4.29812E+18	7686695	385	3	3	1601	5521	7709	576	490234
0	1	4.86057E+18	21560664	37484	2	2	2255103	317	48989	44771	490234
0	1	9.70432E+18	21748480	36759	3	3	4532751	60721	685038	29681	490234
0	1	1.36776E+19	3517124	23778	3	1	1601	2155	1207	1422	490234
0	1	3.28476E+18	20758093	34535	1	1	4532751	77819	266618	222223	490234
0	1	1.01964E+19	21375650	36832	2	1	4688625	202465	457316	429545	490234
0	1	4.20308E+18	4427028	28647	3	1	4532751	720719	3402221	2663964	490234
0	1	4.20308E+18	4428493	28647	2	2	13171922	1493	11658	5668	490234
0	1	5.85475E+17	20945590	35083	2	1	35143	28111	151695	128782	490234
0	1	9.68455E+18	21406020	36943	2	2	4688625	202465	1172072	973354	490234
0	1	4.86057E+18	21560710	37484	2	2	4165614	4107	338524	817824	490234
0	1	1.69552E+19	20730678	34364	2	2	35143	28111	587150	523997	490234
0	1	6.91285E+18	20936539	19186	2	1	34683	61158	81684	373859	490234
0	1	1.18961E+19	10295418	28179	3	2	4532751	720719	2405086	2008317	490234
0	1	6.41431E+18	21183505	35668	2	2	6259	234	15494	1608	490234
0	1	4.86057E+18	21560710	37484	2	2	4165614	4107	338524	572221	490234
0	1	1.16893E+19	21021375	27701	3	2	1601	2155	8580	8736	490234
0	1	1.06646E+19	20620168	30128	2	1	2255103	419	30486	8760	490234
0	1	1.06646E+19	20801912	30128	2	1	13171922	1493	3224	5611	490234
0	1	1.06646E+19	20443036	30128	2	1	4165614	31212	201749	170546	490234
0	1	1.06646E+19	21392028	30128	2	1	4165614	23791	72800	5369	490234
0	1	5.75586E+18	21498278	37333	2	1	12860333	5090	43504	40011	490234
0	1	1.43404E+19	4418786	23808	2	2	12860333	5090	3980	4306	490234
0	1	1.51459E+19	21894794	37932	2	2	34683	138007	531155	425543	490234

UserID Profile Data File

- UserID
- Gender
  - '1' for male, '2' for female, and '0' for unknown
- Age
  - '1' for (0, 12], '2' for (12, 18], '3' for (18, 24], '4' for (24, 30], '5' for (30, 40], and '6' for greater than 40

Methods

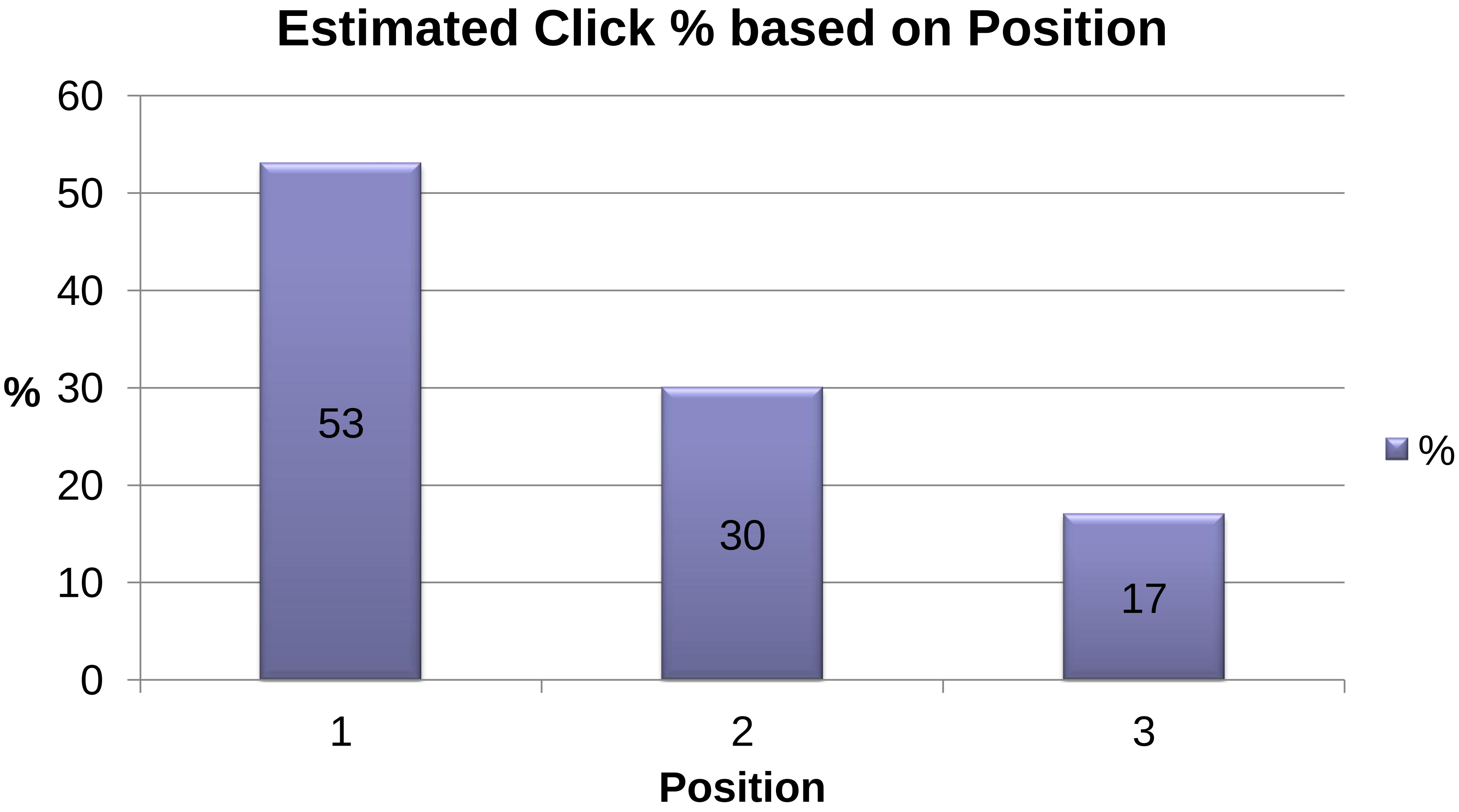
- Python using Hadoop Streaming
- R - Bigmemory package
- doMC package
- Serial Run to verify results

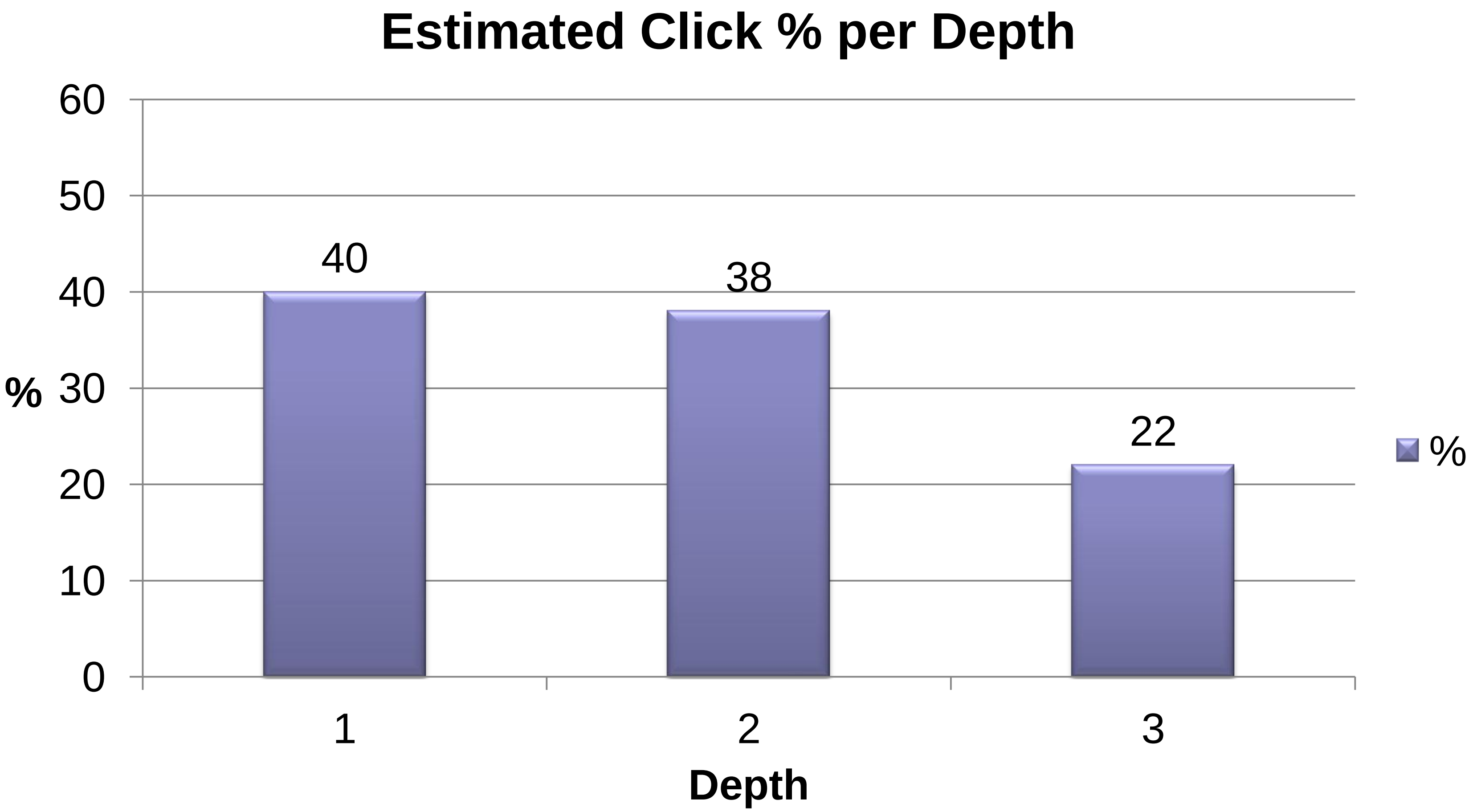
- Each Node Configurations were:
  - 256 gb RAM : 8 cores
  - 16 gb RAM : 8 cores
- Created a new VM and deployed it on each specific node
- Install R and packages, mount hard drives

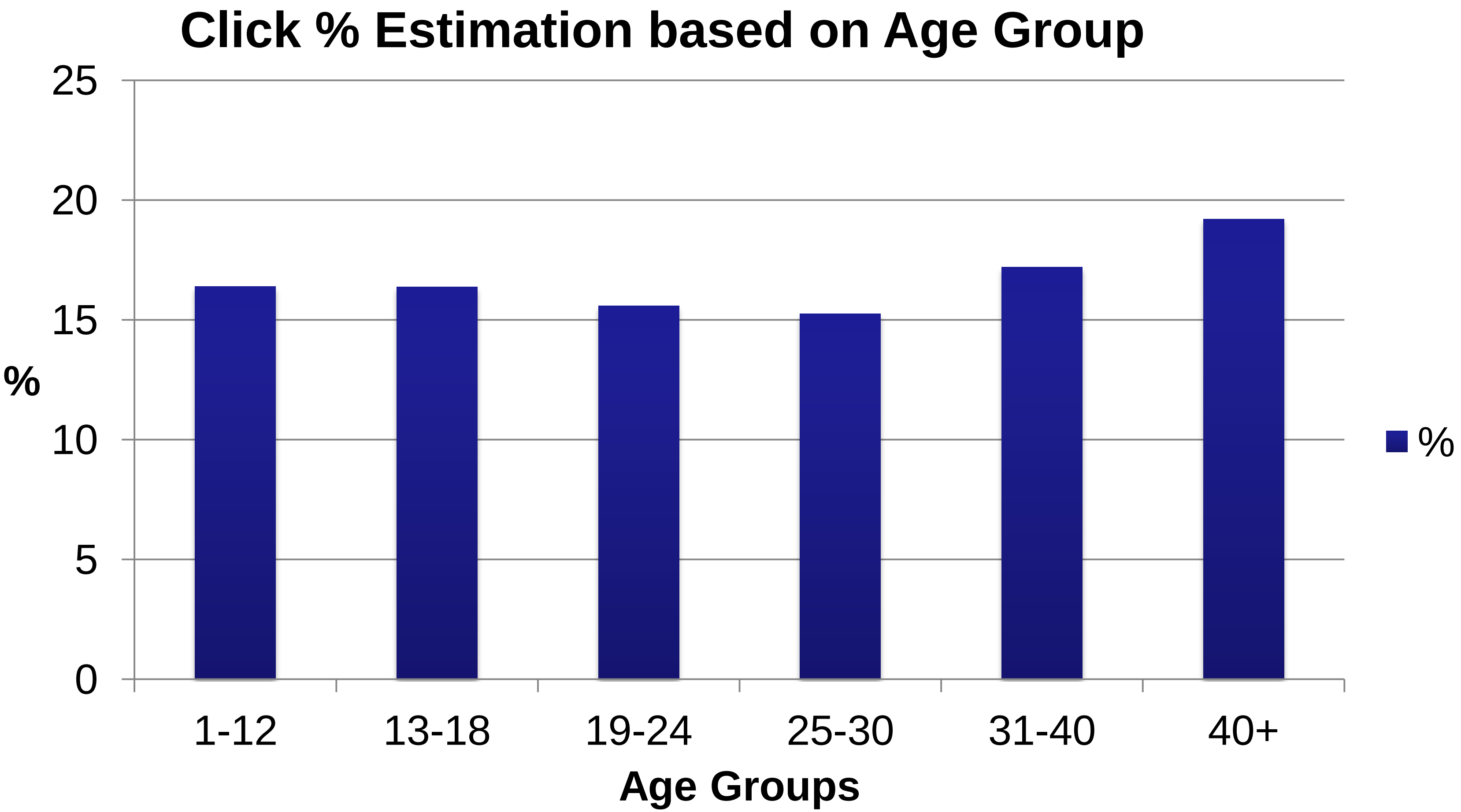
Challenges

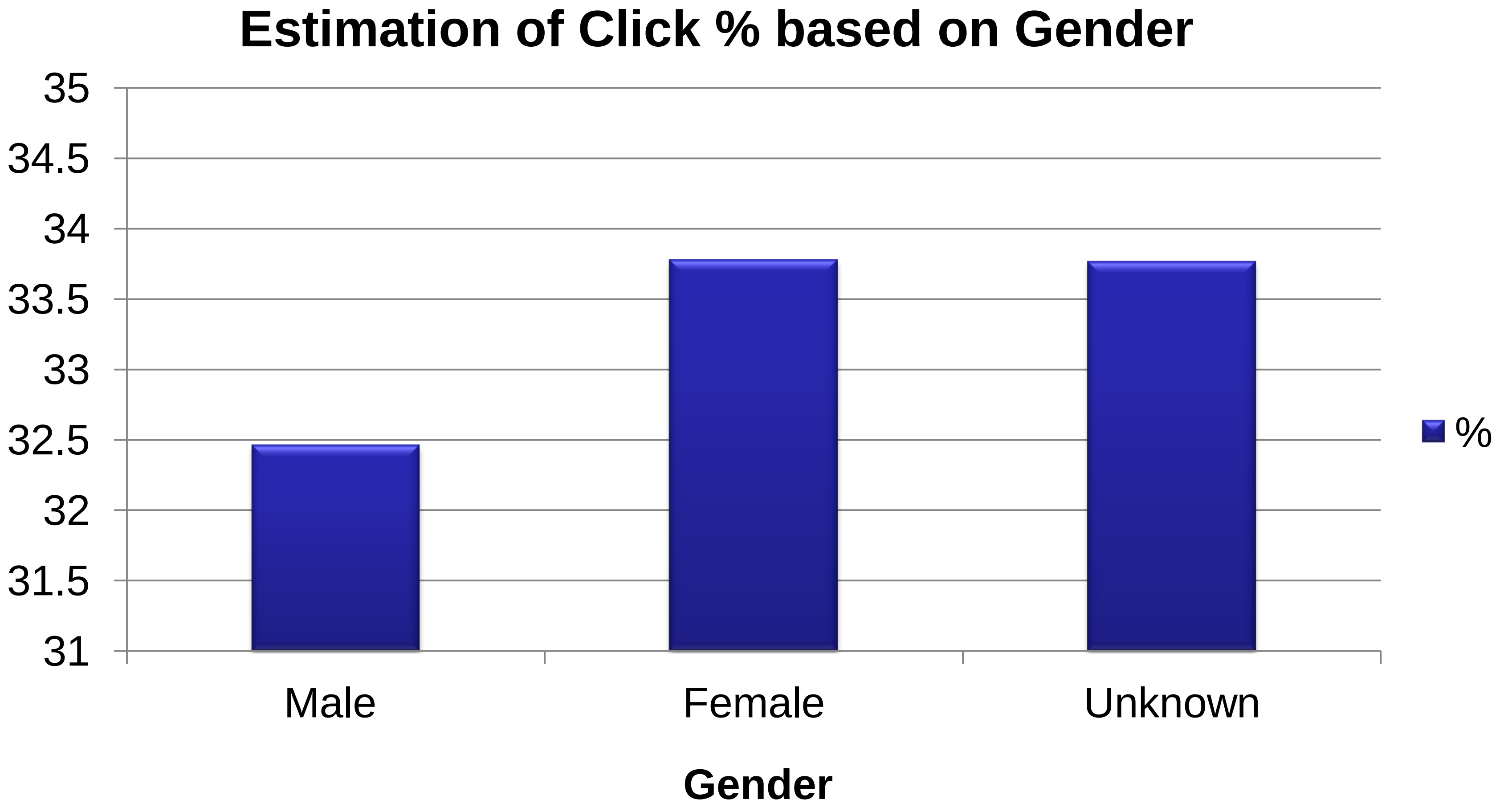
- Brief Documentation on bigmemory
- Merge operation in bigmemory
- Use of proper technique/package to perform Linear Regression
- Ran out of working memory on small machines

Results


Estimated Click % based on Position
60
50
40
%
30
20
10
0
53
30
17
1
2
3
Position
%


Estimated Click % per Depth
60
50
40
%
30
20
10
0
40
38
22
1
2
3
Depth
▨ %


Click % Estimation based on Age Group
25
20
15
%
10
5
0
1-12
13-18
19-24
25-30
31-40
40+
Age Groups
■ %


Estimation of Click % based on Gender
35
34.5
34
33.5
33
32.5
32
31.5
31
Male
Female
Unknown
Gender
%

Linear Regression

Sample size = 149,639,105 = 149 million records

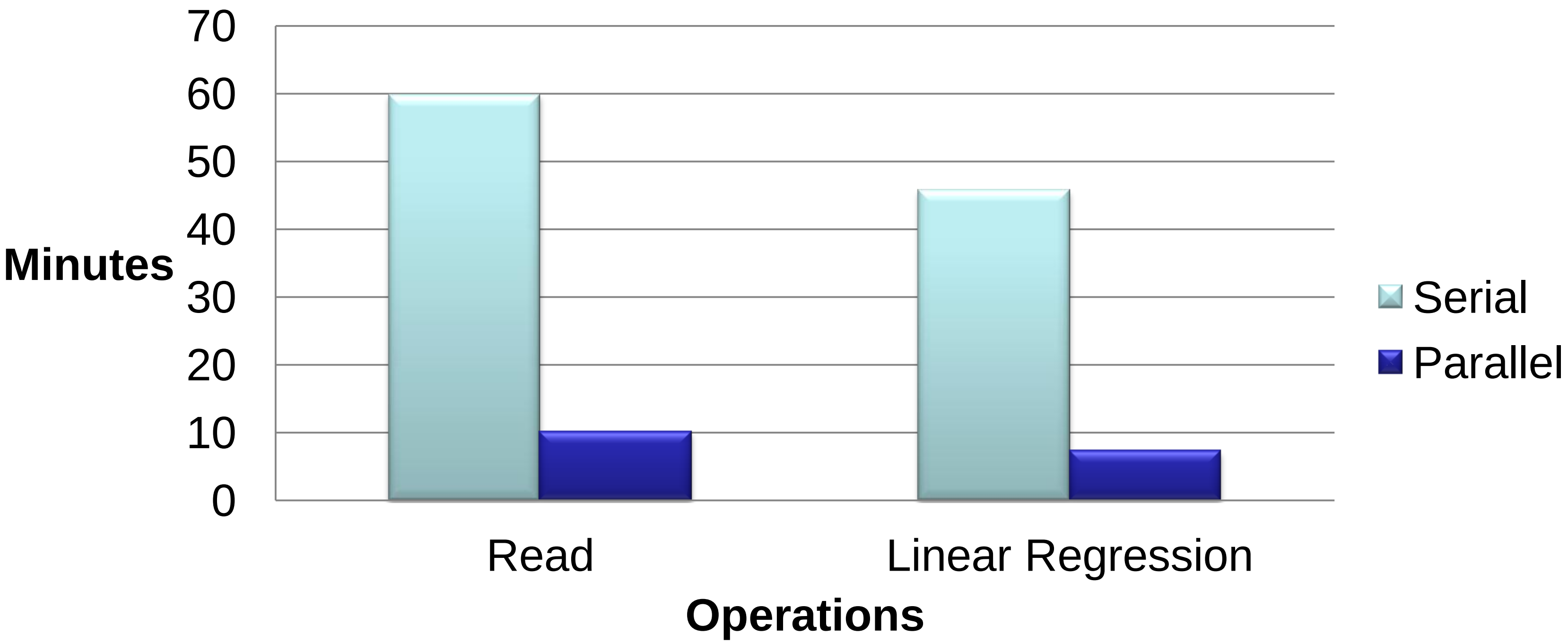
	Coef	95%	CI	P
(Intercept)	0.0988	0.0959	0.1016	0.0000
Depth	-0.0003	-0.0019	0.0013	0.7191
Position	-0.0291	-0.0308	-0.0273	0.0000

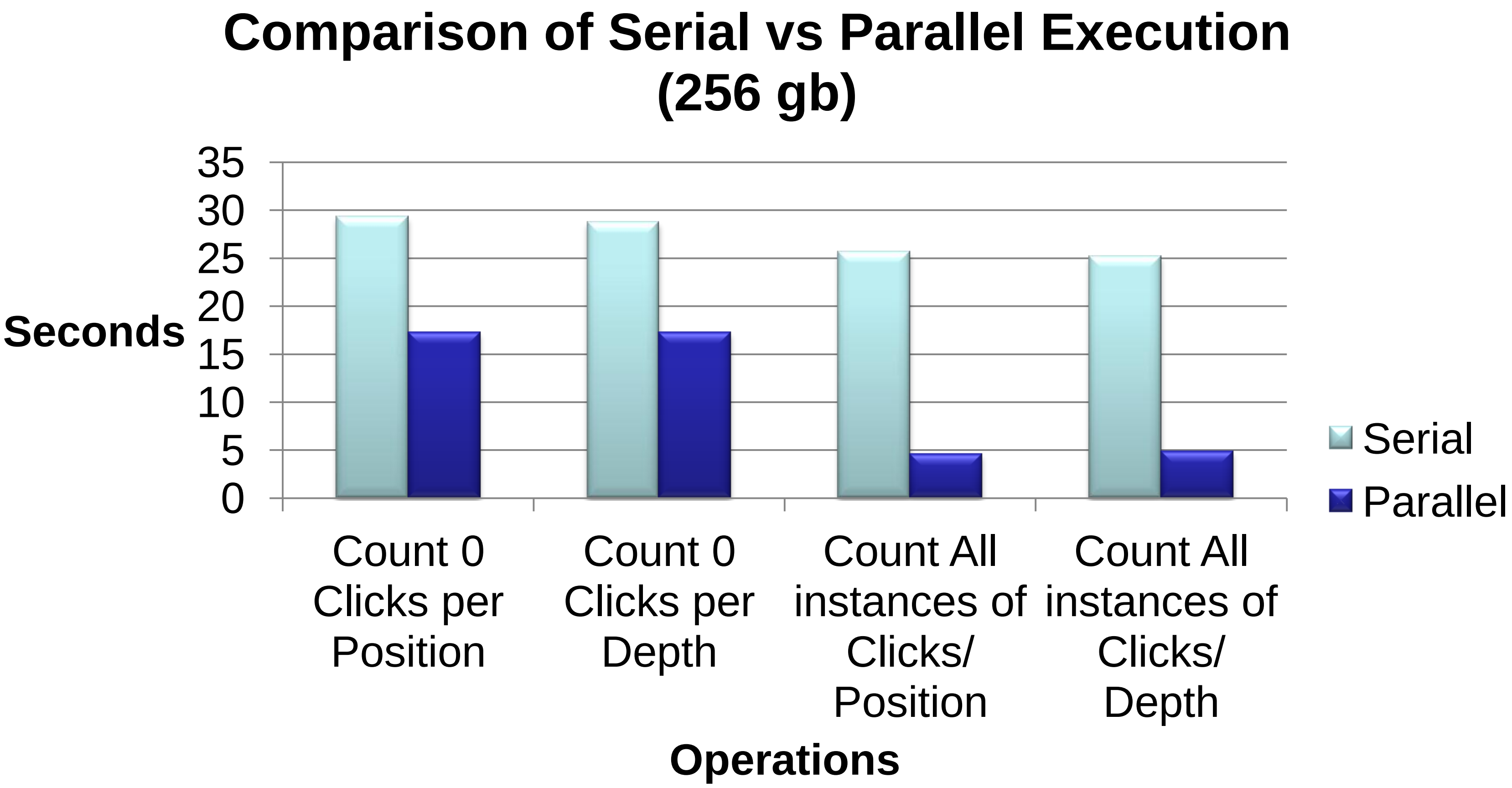
Performance Benchmarks

- Compute Node Hardware Specifications

Node	Processor	Cores	Memory
Node 1188	Intel Xeon L5420 @2.5GHz	8	16 gb
Nodelm03	Intel Xeon 7542 @ 2.66 GHz	8	256 gb

Comparison of Serial vs Parallel Execution (256 gb)




Comparison of Serial vs Parallel Execution
(256 gb)
Seconds
35
30
25
20
15
10
5
0
Count 0 Clicks per Position
Count 0 Clicks per Depth
Count All instances of Clicks/ Position
Count All instances of Clicks/ Depth
Operations
Serial
Parallel

	Serial	Parallel	Parallel
Operation	(256 gb)	(256 gb)	(16 gb)
Read	59.88 mins	10.18 mins	14.70 mins
Count 0 Clicks/ Position	29.30 secs	17.30 secs	25.20 secs
Count 0 Clicks / Depth	28.75 secs	17.30 secs	27.80 secs
Count All Instances/ Position	25.67 secs	4.6 secs	14.30 secs
Count All Instances / Depth	25.23 secs	4.9 secs	10.80 secs
Linear Regression	45.80 mins	7.30 mins	

Questions?