DSC 102
Systems for Scalable Analytics

Arun Kumar

Surprise Quiz 5
Which of the following gradient methods does not require a shuffle (to randomize the data ordering)?

A  BGD
B  SGD with batch size of 1
C  SGD with batch size > 1
D  All of A, B, C
E  None of A, B, C
Surprise Quiz

Which of the following linear algebra/matrix operations does not have a filescan access pattern in general?

A  Frobenius (L2) norm
B  L1 norm
C  Matrix multiplication
D  Max norm
E  None of the above
Which of the following data layout strategies for structured data can *not* be used with BSP at all?

A  Row-oriented
B  Column-oriented
C  Tile-oriented
D  All of A, B, C
E  None of A, B, C
Surprise Quiz

What is a key efficiency advantage of Parameter Server over model averaging for distributed SGD?

A Scales to multi-node clusters
B Converges for GLMs
C Asynchronous updates
D Parallelizes gradient computation
E All of the above
Which of the following is often a major practical challenge for data scientists in the data sourcing stage?

A  Messiness of data  
B  Access control restrictions  
C  Heterogeneity of data  
D  Scale of data  
E  All of the above
Which of the following applications is typically likely to need extra labeling effort to use supervised ML?

A  E-retail recommendation systems
B  Enterprise sales forecasting
C  Computational advertising
D  Radiology in healthcare
E  None of the above
At what stage of data sourcing is entity matching typically handled in practice?

A  Acquisition
B  Data re-organization
C  Data cleaning
D  Data labeling
E  Feature engineering
Answers in order of questions

A, C, E, C, E, D, C