Question 1: Compression . . . . . [400 points]

(i) **[10 points] Bottlenecks:**
Distinguish between I/O bound and CPU bound programs.

(ii) **[10 points] Bottlenecks:**
Explain the CPU-DRAM performance gap. Is it growing or shrinking?

(iii) **[10 points] Compression:**
Explain the trade-off between decompression speed and compression ratio.

(iv) **[10 points] Compression:**
Why does a disk-centric DBMS optimize for compression ratio? Why does an in-memory DBMS optimize for decompression speed?

(v) **[10 points] Compression:**
Distinguish between DRAM bandwidth and disk bandwidth.

(vi) **[10 points] Real-World Data Characteristics:**
Define skew in real-world distribution of data. Illustrate with an example.

(vii) **[10 points] Real-World Data Characteristics:**
Define correlation in real-world distribution of data. Illustrate with an example.

(viii) **[10 points] Real-World Data Characteristics:**
How can a compression scheme leverage: (1) skew, and (2) correlation?

(ix) **[10 points] Database Compression:**
Why do we seek for a compression scheme to produce fixed-length values?

(x) **[10 points] Database Compression:**
Distinguish between early and late materialization from a query execution standpoint.

(xi) **[10 points] Database Compression:**
Distinguish between lossy and lossless compression.

(xii) **[20 points] Data Skipping:**
Define approximate query processing. Explain with an illustrative query.

(xiii) **[20 points] Data Skipping:**
Define zone maps. Explain with an illustrative query.

(xiv) **[10 points] Compression Granularity:**
Is compression better suited for NSM or DSM storage model? Justify your answer.

(xv) **[10 points] Naive Compression:**
What is the entropy of a data distribution? How is it related to thermodynamics?

(xvi) **[10 points] Naive Compression:**
What is entropy encoding?

(xvii) **[10 points] Naive Compression:**
What is dictionary encoding?

Question 1 continues…

(xviii) **[10 points] Naive Compression:**
How is entropy encoding related to dictionary encoding?

(xix) **[10 points] InnoDB Compression:**
Briefly explain how compression is done in MySQL InnoDB storage engine?

(xx) **[10 points] Naive Compression:**
Illustrate how we could execute point queries on compressed data.

(xxi) **[10 points] Naive Compression:**
Illustrate how we could execute range queries on compressed data.

(xxii) **[10 points] Compression Schemes:**
Define Null Suppression. Illustrate with an example.

(xxiii) **[10 points] Compression Schemes:**
Define Run Length Encoding. Illustrate with an example.

(xxiv) **[10 points] Compression Schemes:**
Define Bitmap Encoding. Illustrate with an example.

(xxv) **[10 points] Compression Schemes:**
Why is the efficacy of bitmap encoding dependent on the cardinality of the attribute?

(xxvi) **[10 points] Compression Schemes:**
Is it effective to compress zip codes using bitmap encoding?

(xxvii) **[10 points] Compression Schemes:**
Is it effective to compress zip codes using bitmap encoding?

(xxviii) **[10 points] Compression Schemes:**
How is Byte-Aligned Bitmap Code (BBC) "byte-aligned"?

(xxix) **[10 points] Compression Schemes:**
Do these schemes support random access to a given value? Justify your answer.

(xxx) **[10 points] Compression Schemes:**
Define Delta Encoding. Illustrate with an example.

(xxxi) **[10 points] Compression Schemes:**
Define Incremental Encoding. Illustrate with an example.

(xxxii) **[10 points] Compression Schemes:**
Define Mostly Encoding. Illustrate with an example.

(xxxiii) **[10 points] Compression Schemes:**
Illustrate how you could combine two compression schemes to get an even better compression ratio.

(xxxiv) **[10 points] Dictionary Compression:**
Why does this scheme not require pre-sorting?

(xxxv) **[10 points] Dictionary Compression:**
Why do we need to re-encode existing values with incremental construction?

Question 1 continues…

(xxxvi) **[10 points] Order-Preserving Encoding:**
Define an order-preserving encoding. Illustrate with an example.

(xxxvii) **[10 points] Order-Preserving Encoding:**
Can we short-circuit query execution using a dictionary?s Illustrate with an example.

(xxxviii) **[10 points] Dictionary Data Structures:**
Why is a B+Tree better suited for building a dictionary?