ID2221 - Data-Intensive Computing (Dec. 19, 2024)

Examiner: Amir H. Payberah (phone num: 072-55 44 011)
No aids are allowed
The exam consists of 20 multiple-choice questions, each with only one correct answer, and each question has one point. The final grading for the exam is as follows:
A: $x \geq 18$, **B:** $18 < x \leq 16$, **C:** $16 < x \leq 14$, **D:** $14 < x \leq 12$, **E:** $12 < x \leq 10$, **F:** $x < 10$

---

1. How does GFS distinguish between up-to-date and stale replicas?

   (a) By periodically comparing data across all chunkservers.
   (b) By assigning version numbers to chunks and updating them when leases are granted.
   (c) By storing replicas in different racks for cross-verification.
   (d) By maintaining logs of all write operations.

---

2. What is the purpose of the Chubby lock service in BigTable?

   (a) To replicate data across tablet servers.
   (b) To manage read and write operations.
   (c) To store BigTable schema information and access control lists.
   (d) To cache tablet locations.

---

3. What is a primary benefit of the column-oriented data model compared to the row-oriented model?

   (a) Better integration with SQL queries.
   (b) Improved storage of hierarchical data.
   (c) Faster operations on specific columns of data.
   (d) Simplified schema management.

---

4. When a new node is added to a Cassandra cluster that uses consistent hashing, what happens to the existing data?

   (a) All existing data is rehashed and redistributed across all nodes, including the new node.
   (b) The new node takes over responsibility for a portion of the data, but no data is moved from existing nodes.
   (c) Only the data that falls within the new node's hash range is moved from existing nodes to the new node.
   (d) The new node remains idle until a sufficient amount of new data is inserted to warrant its use.

5. What trade-off does BigTable make by prioritizing strong consistency over availability?

   (a) Temporary unavailability of data when a tablet server fails.
   (b) Increased redundancy.
   (c) Reduced read/write latency.
   (d) Simplified replication process.

---

6. What is the role of the shuffle and sort phase in MapReduce?

   (a) To execute the map tasks in parallel.
   (b) To divide the input data into smaller splits.
   (c) To replicate the reduce tasks across multiple nodes.
   (d) To group and consolidate intermediate key-value pairs by keys.

---

7. In a reduce-side join in MapReduce, what must be ensured before performing the join operation?

   (a) Data is sorted and grouped by the join key.
   (b) Both datasets are partitioned into equal-sized splits.
   (c) One of the datasets is small enough to fit in memory.
   (d) Intermediate data is written to the master node.

---

8. Which of the following best describes the role of the lineage graph in Spark?

   (a) It represents the sequence of RDD operations for fault tolerance.
   (b) It is used to schedule tasks across executors.
   (c) It determines the partitioning of data across the cluster.
   (d) It handles the caching of RDDs in memory.

---

9. What is the primary advantage of caching an RDD in Spark?

   (a) It ensures the RDD is distributed across all executors.
   (b) It allows the RDD to be checkpointed to disk.
   (c) It reduces the number of partitions in the RDD.
   (d) It avoids re-evaluating the RDD for subsequent actions.

---

10. What is the output of the `groupBy` operation in a Spark DataFrame?

    (a) A new DataFrame with grouped data ready for aggregation.
    (b) A Dataset containing only distinct rows.
    (c) A DataFrame with columns renamed based on the grouping keys.
    (d) A new RDD with keys representing grouping attributes.

11. What is the key advantage of using Datasets over DataFrames?

    (a) Datasets allow untyped transformations.
    (b) Datasets automatically cache all transformations.
    (c) Datasets provide compile-time type safety.
    (d) All of the above.

---

12. What is the purpose of watermarks in stream processing?

    (a) To synchronize event time and processing time.
    (b) To buffer data for micro-batch processing.
    (c) To identify partitions within a topic.
    (d) To provide a threshold indicating how long the system waits for late events.

---

13. Which of the following best describes a sliding window in stream processing?

    (a) A window that retains a fixed number of tuples before discarding old ones.
    (b) A window that supports batch operations by evicting all tuples after processing.
    (c) A window that allows overlapping buffers for incremental processing.
    (d) A window that aggregates data over a fixed period.

---

14. In Kafka, what is the significance of the offset associated with each message?

    (a) It uniquely identifies the position of a message within a partition.
    (b) It ensures at-least-once delivery of messages.
    (c) It guarantees the total ordering of messages across partitions.
    (d) It enables the replication of messages across brokers.

---

15. Why is checkpointing required for stateful operations in Spark Streaming?

    (a) To persist data to external storage for later analysis.
    (b) To store RDD transformations for fault tolerance.
    (c) To save intermediate states and ensure recovery in case of failures.
    (d) To enable window-based transformations.

---

16. What is the main difference between Spark Streaming and Structured Streaming?

    (a) Spark Streaming uses DStreams, while Structured Streaming uses continuous tables.
    (b) Spark Streaming supports window operations, while Structured Streaming does not.
    (c) Spark Streaming processes data in real-time, while Structured Streaming processes in batches.
    (d) Spark Streaming handles late data natively, while Structured Streaming does not.

17. What is the main challenge with using edge-cut partitioning for natural graphs?

    (a) It does not balance the number of vertices across partitions.
    (b) It performs poorly on skewed power-law degree distributions.
    (c) It requires duplicating all edges across partitions.
    (d) It is incompatible with iterative graph algorithms.

---

18. Why is asynchronous computation in GraphLab often faster than synchronous models like Pregel?

    (a) It processes vertices sequentially rather than in parallel.
    (b) It requires fewer iterations to converge on large graphs.
    (c) It avoids partitioning the graph altogether
    (d) It eliminates the need for global synchronization between supersteps.

---

19. What is a key limitation of max-min fairness in resource allocation?

    (a) It does not guarantee optimal resource utilization.
    (b) It cannot handle scenarios with heterogeneous resources.
    (c) It prioritizes certain users over others based on task requirements.
    (d) It leads to frequent re-allocation of resources.

---

20. How does Delta Lake ensure data reliability in its tables?

    (a) By replicating data across multiple storage layers.
    (b) By caching all operations in memory.
    (c) By restricting updates to batch processes only.
    (d) Through a DeltaLog that tracks changes and ensures consistency.