## Data-Juicer: Partitioning, Checkpointing & Event Logging System **Input Dataset** Configuration **Work Directory** • JSONL/Parquet Files • YAML Config Files Partitions • Pipeline Operations Large Datasets Checkpoints • Remote URLs System Settings • Event Logs EnhancedPartitionedRayExecutor Partitioning Engine **DatasetBuilder EventLogger** • Track All Events Load Dataset • Split Dataset • Format Detection • Real-time Monitoring • Schema Inference • Metadata Generation Performance Metrics Checkpoint Manager **Ray Cluster Result Merger** Save States • Distribute • Combine Partitions • Load States • Parallel Execution Validate Results Cleanup • Fault Handling • Export Dataset **Performance Output Dataset Event Logs** Report • JSONL Format Processed Data • Timing Analysis Validated Results Rotated Logs Resource Usage Optimized Format Compressed Storage • Bottleneck Detection