Execution of analytic frameworks on sample data sets is the current trend in response to increasing data size and demand for real-time analysis. Additionally, high-performance, energy-efficient Solid-State Drives (SSD) arrays are the primary storage subsystem for parallel data analysis systems. To exploit the benefits of SSD arrays when executing applications, several key areas must be considered. First, due to logical to physical address translation, random data choice in data sampling jobs can cause unbalanced workloads among SSDs in the array. Second, after the data choice, existing task schedulers in data analysis frameworks can introduce non-negligible resource contentions resulting from the suboptimal Input/Output (I/O). The performance of SSDs is unpredictable because of their varying maintenance costs at runtime, which renders them hard to be managed by the scheduler. With the trend towards sample set data analytics and the use of SSDs, it is increasingly important to ensure balanced workloads and minimize resource contentions. Without addressing these areas, sample-set data analytics on SSDs will continue to suffer from performance inefficiencies. Without addressing these areas, sample-set data analytics on SSDs will continue to suffer from performance inefficiencies. This dissertation presents ApproxSSD to perform on-disk layout-aware data sampling on SSD arrays. This proposed framework leverages data selection and task scheduling to improve the performance of many applications. ApproxSSD decouples I/O from the computation in task execution. This avoids potential I/O contentions and suboptimal workload balances.