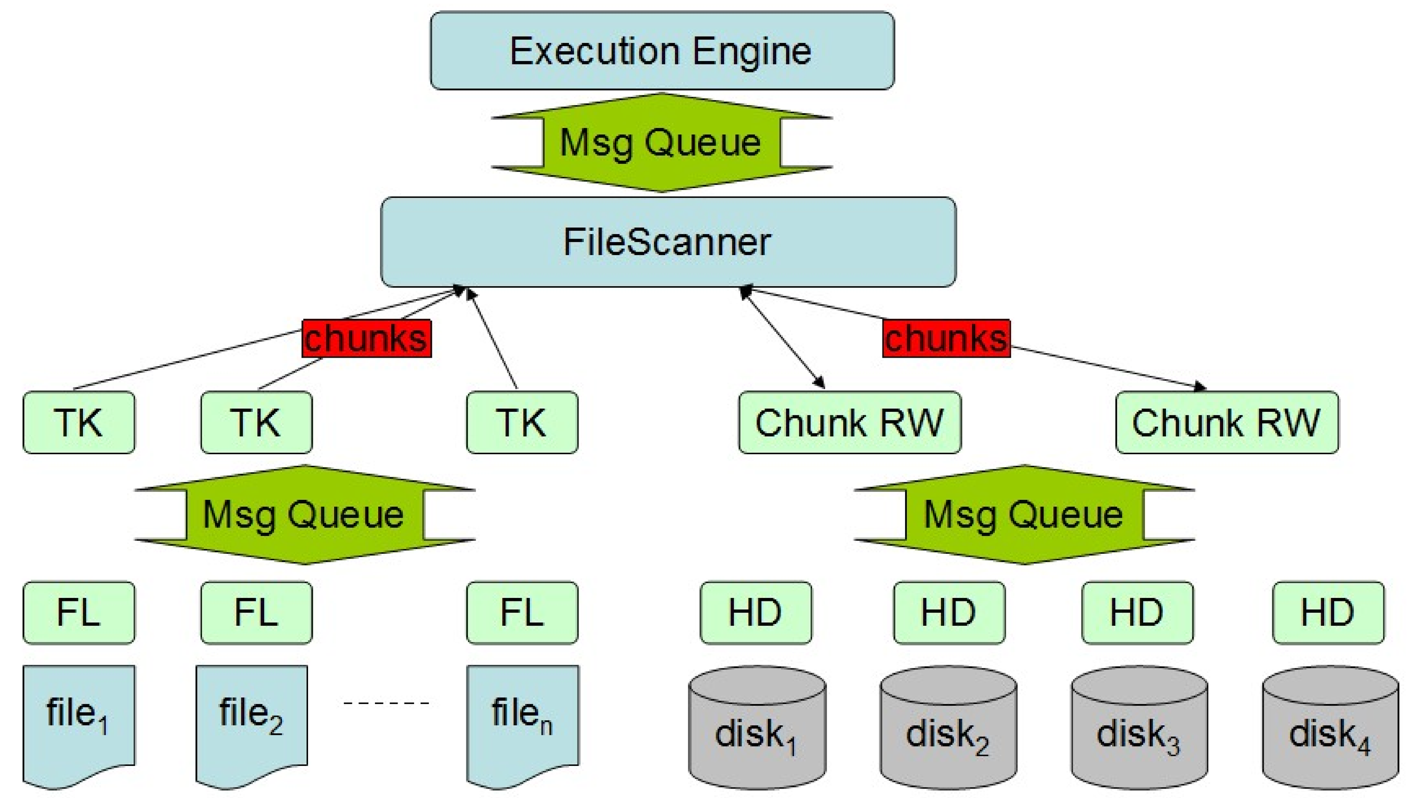
Running Queries From Raw Files

Yu Cheng, Florin Rusu
Electrical Engineering and Computer Science, University of California, Merced
ycheng4@ucmerced.edu, frusu@ucmerced.edu

Storage Manager Architecture

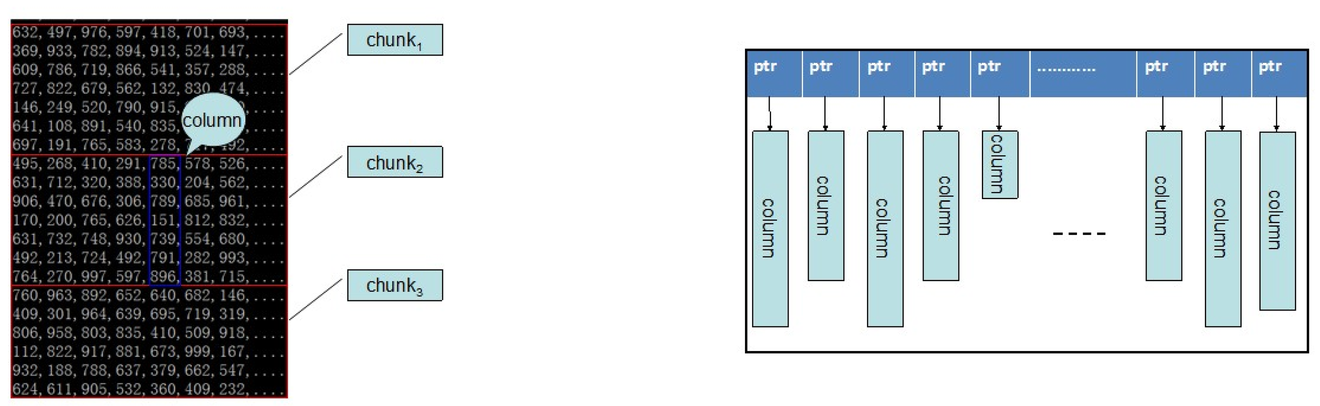
Storage Manager is responsible for reading/writing data from/to database tables and/or external files



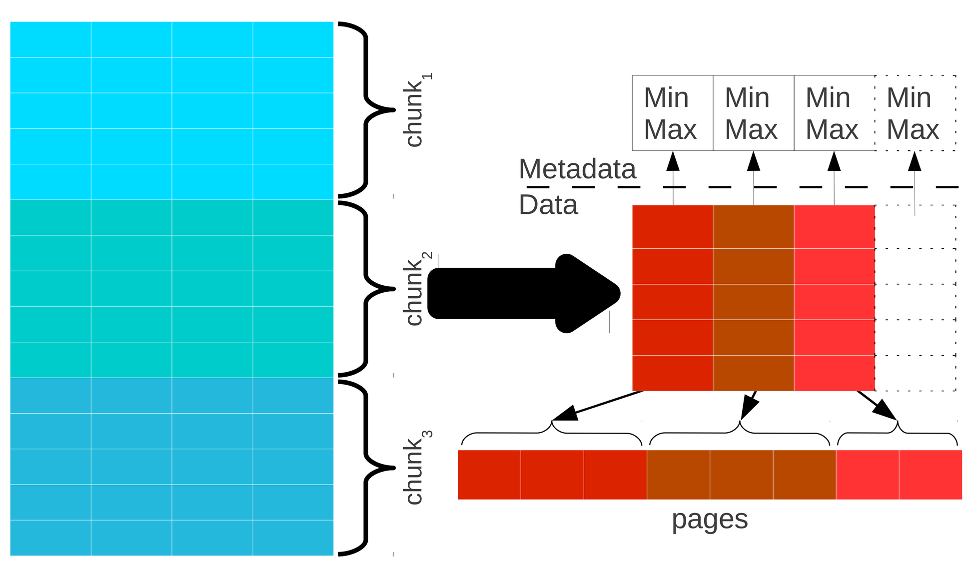
- Heterogeneous data source: Data retrieving from database tables and raw files
- Parallel processing: Split time-consuming stages into several components and apply pipelining mechanisms
- Architecture independence: multi-thread (shared memory and shared disk) and inter-node (shared nothing) parallelism

Chunk Structure

Chunks are both the I/O and processing unit



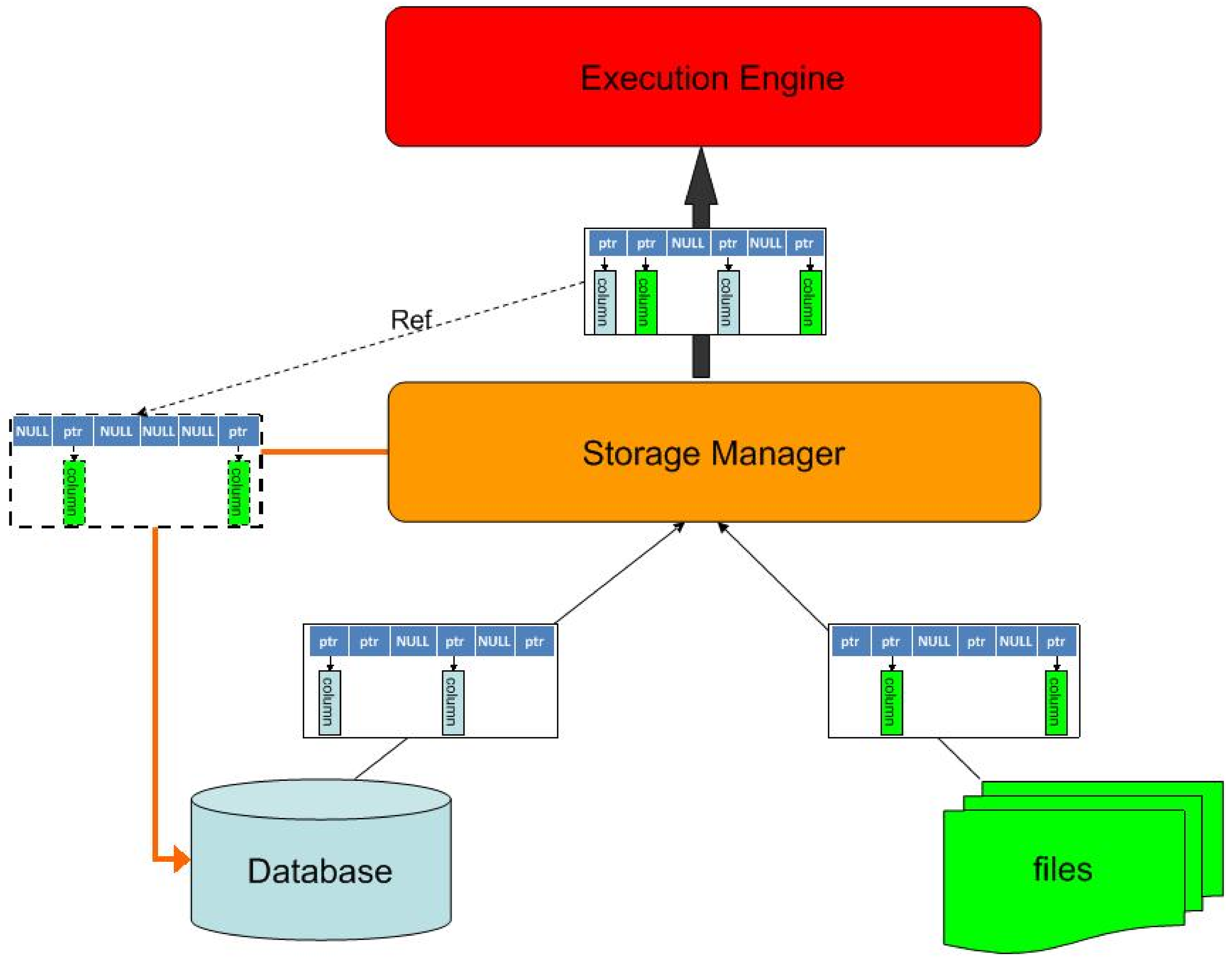
- Structure: a chunk is composed of columns connected by address pointers



- Metadata: each chunk contains the minimum and maximum values for each column

On-the-fly Loading

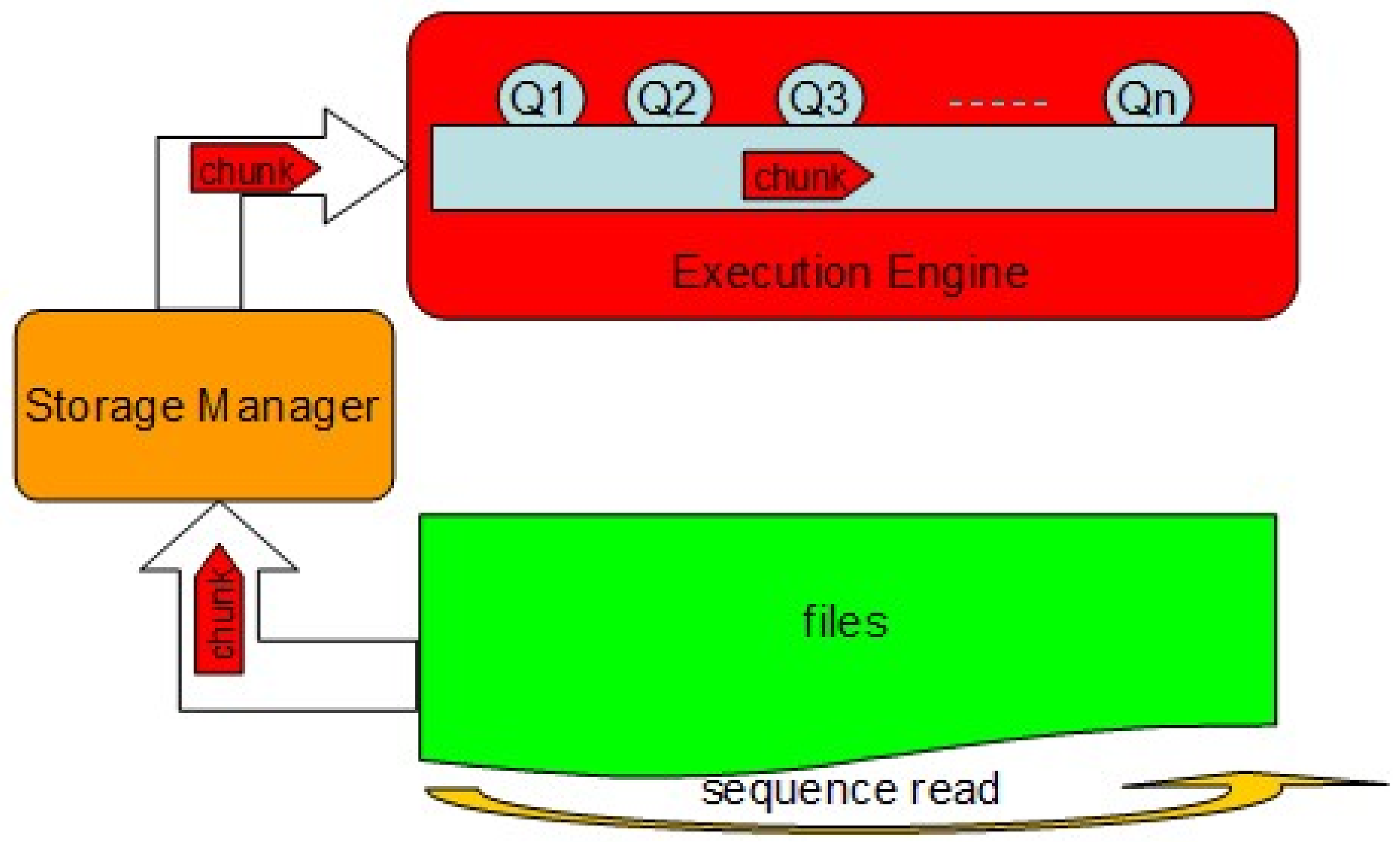
Execute queries immediately from raw files without previous loading



- Start-up: each query can be executed directly from the raw file
- Multi-source support: query data can be from database, raw files, or both
- On-the-fly load: load data during query processing

Execute Batch Queries

System supports executing multiple queries concurrently



- Maximize utilization: make full use of each reading procedure

Experiments and Results

System: Ubuntu SMP 11.04 with Linux kernel 2.6.38-14, 16 cores, 16GB of RAM, and 4 1TB disks.

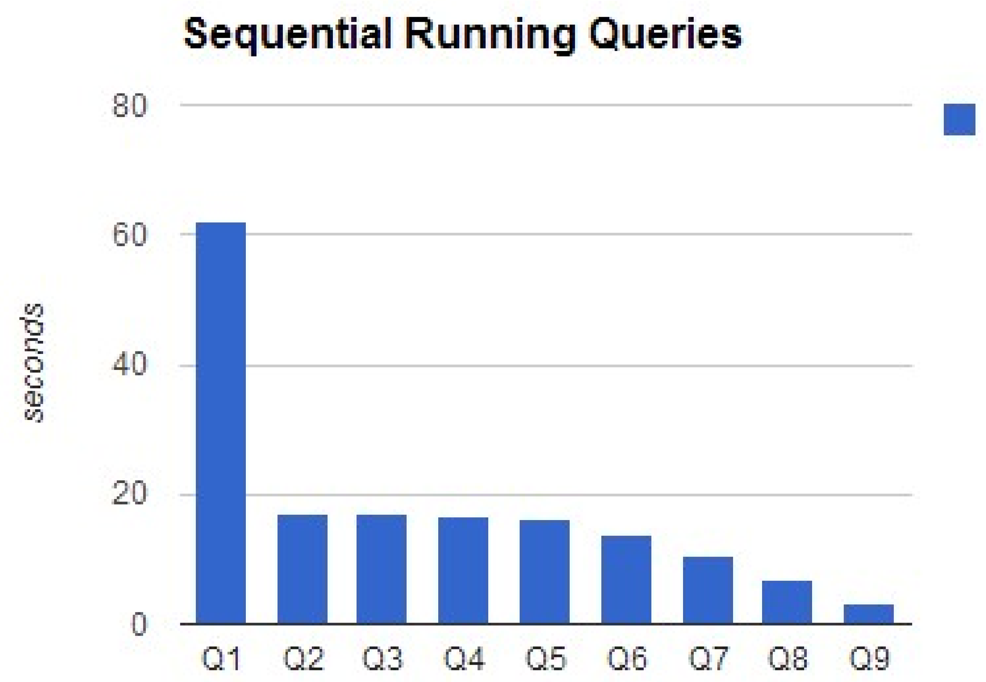
Dataset: raw file with $7.5 * 10^6$ tuples (11GB). Each tuple contains 150 attributes with random integers uniformly distributed in $[0 - 10^9]$

Task: sum of a subset of attributes with different selection conditions

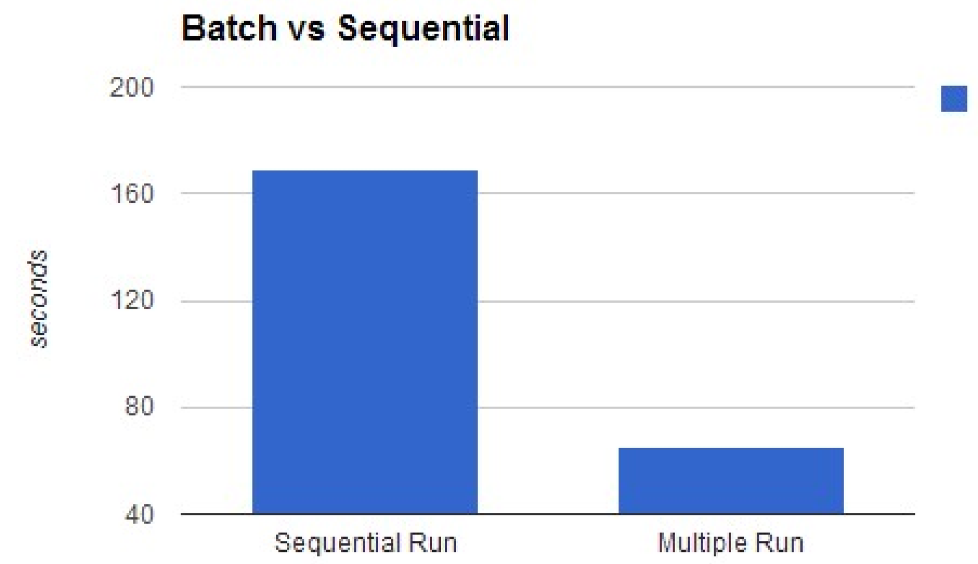
Query	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
No. of attributes	150	150	150	150	150	120	90	60	30
Query selectivity (%)	100	80	60	40	20	100	100	100	100

Results

- Execute queries sequentially



- Sequential vs. batch execution



- Raw file execution vs. on-the-fly loading vs. database