SwiftTuna:
Incrementally Exploring Large-scale Multidimensional Data

Jaemin Jo jmjo@hcil.snu.ac.kr
Wonjae Kim wjkim@hcil.snu.ac.kr
Seunghoon Yoo shyoo@hcil.snu.ac.kr
Bohyoung Kim bkim@hufs.ac.kr
Jinwook Seo jseo@snu.ac.kr

Seoul National University
Dept. of Computer Science & Engineering

Summary

Challenges in low-latency data exploration of large-scale data

- Precomputed data structures (e.g., data cubes) have been often used
- Required a large amount of memory (e.g., limited # of dimensions)
- Only targeted single machine scenarios

Rapid and incremental exploration without precomputation

- Incremental data processing for responsiveness
- Visualizations designed for scalability
- Uncertainty visualizations for approximate queries

Bringing modern cluster computing technologies to InfoVis

- Exploit in-memory computing engine (i.e. Apache Spark)

Design Consideration

Process results incrementally while estimating the final results

- Adopted gradient plots to visualize the uncertainty of partial results
- 95% confidence intervals of counts and means

Enable flexible scheduling of queries

- Pause or stop queries in real time if partial results are enough

Scalability in visualizations

- Binned plots with the Focus+Context techniques
- Designed tailed charts to summarize many categories on the x-axis

Provide low-fidelity feedback promptly

- Based on a small sample from the data (i.e., 0.001% of entries)

Performance Benchmark

Used Criteo's Terabyte Click Logs dataset

- 1.03 TB csv, 4.3B entries and 40 dimensions

16 r3.8xlarge instances on Amazon Web Services (AWS)

- Intel E5-2670 v2 (32 vCPUs), 244 GB of memory, and 2 * 320 GB SSD

Measured mean interval between two successive responses

- 240 blocks (1.75M rows per block) and 2,400 blocks (17.5M rows per block)

Type	Range or Cardinality	2,400 Blocks (s)	240 Blocks (s)
Binned Histogram	0 – 35M	1.91±0.84	3.54±1.58
Density Plot	0 – 746K, 0 – 35M	1.88±0.61	3.46±1.05
Frequency Histogram	20K	2.85±0.78	3.93±1.31
Pivot Dot Plots (MEAN)	0 – 35M, 7.4K	2.53±1.21	3.88±0.93

Each incremental process on a block took approx. 2 seconds

Trade-off between responsiveness and throughput

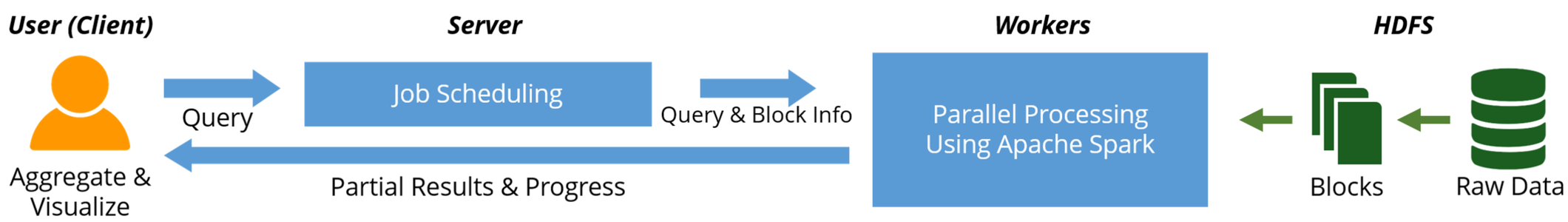
- Smaller blocks ⟶ better responsiveness, larger blocks ⟶ better throughput
- Find the optimum number and size of blocks

Conclusion & Future Work

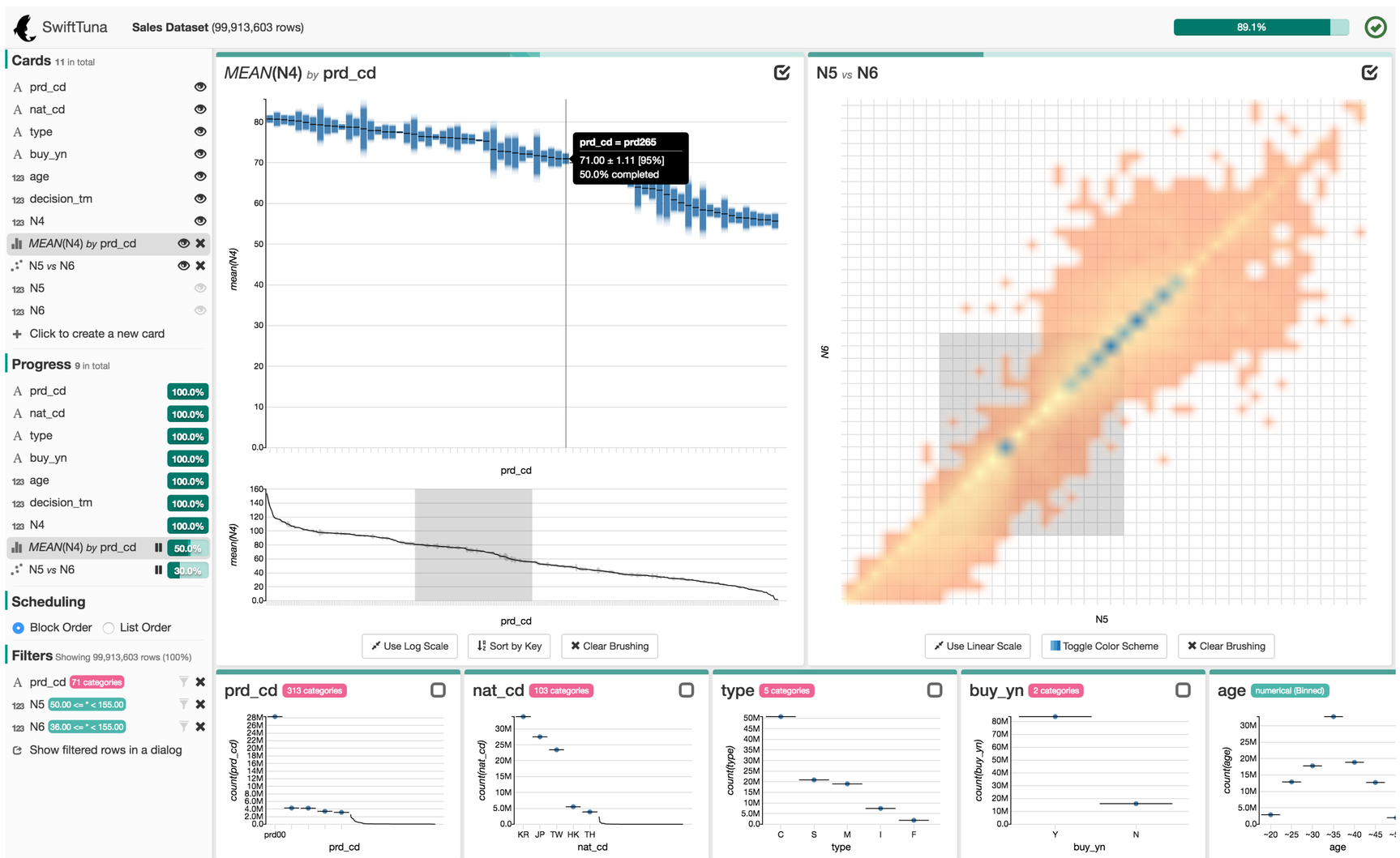
Proposed an interactive system for fluent exploration of large-scale multidimensional data

- Harmony between information visualization and distributed computing

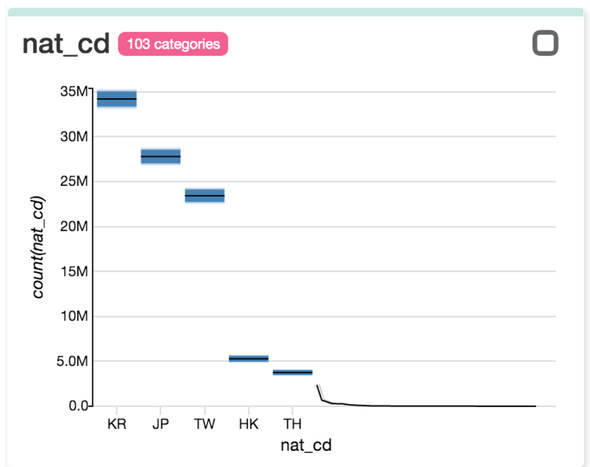
Extend the system to a general platform for incremental visual analytics



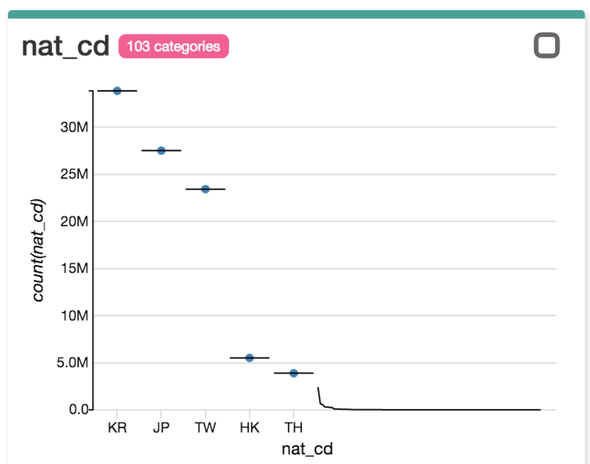
Interface Design



The main interface of SwiftTuna



A tailed gradient plot is showing confidence intervals



A tailed dot plot