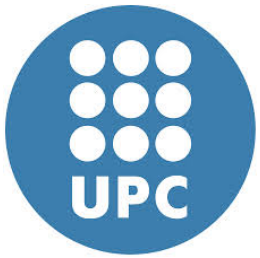
Sparksee Feedback and the future of SNB

Arnau Prat[1]

[1]Sparsity-Technologies
DAMA-UPC

LDBC TUC Meeting Athens
November 2014



SNB in Sparksee

- Implemented the 14 SNB interactive workload look up queries
  - Validated using the official validation dataset
  - Tested against SF1 for preliminar resuls

SNB in Sparksee

- Implemented the 14 SNB interactive workload look up queries
  - Validated using the official validation dataset
  - Tested against SF1 for preliminar resuls
- Use SNB to profile Sparksee

SNB in Sparksee

- Implemented the 14 SNB interactive workload look up queries
  - Validated using the official validation dataset
  - Tested against SF1 for preliminar resuls
- Use SNB to profile Sparksee
- Our short term targets are SF100 and SF300 with updates
  - Impact of going out of core
  - Multicore scalability
  - Java vs C++ implementations
  - Critical bottlenecks, API improvements (Language?)

SNB in Sparksee

- Implemented the 14 SNB interactive workload look up queries
  - Validated using the official validation dataset
  - Tested against SF1 for preliminar resuls
- Use SNB to profile Sparksee
- Our short term targets are SF100 and SF300 with updates
  - Impact of going out of core
  - Multicore scalability
  - Java vs C++ implementations
  - Critical bottlenecks, API improvements (Language?)
- Integrate SNB into our development pipeline

SNB in Sparksee - Things learned so far

- Query validation is painful
  - Better result mismatch reporting would be appreciated

SNB in Sparksee - Things learned so far

- Query validation is painful
  - Better result mismatch reporting would be appreciated
- Expensive queries if not properly implemented
  - Avoid intermediate results materialization
  - Exploration order affects a lot

SNB in Sparksee - Things learned so far

- Query validation is painful
  - Better result mismatch reporting would be appreciated
- Expensive queries if not properly implemented
  - Avoid intermediate results materialization
  - Exploration order affects a lot
- Extending the schema to reduce the search space
  - Found that most users belonging to forums do not post anything
  - Is this as intended? The same for larger SF?

SNB in Sparksee - Things learned so far

- Exploiting correlations with Sparksee internal ids can be beneficial
  - Many queries ask for results sorted by date
  - Sparksee "Objects" collection is always sorted by id
  - We can avoid performing sorts

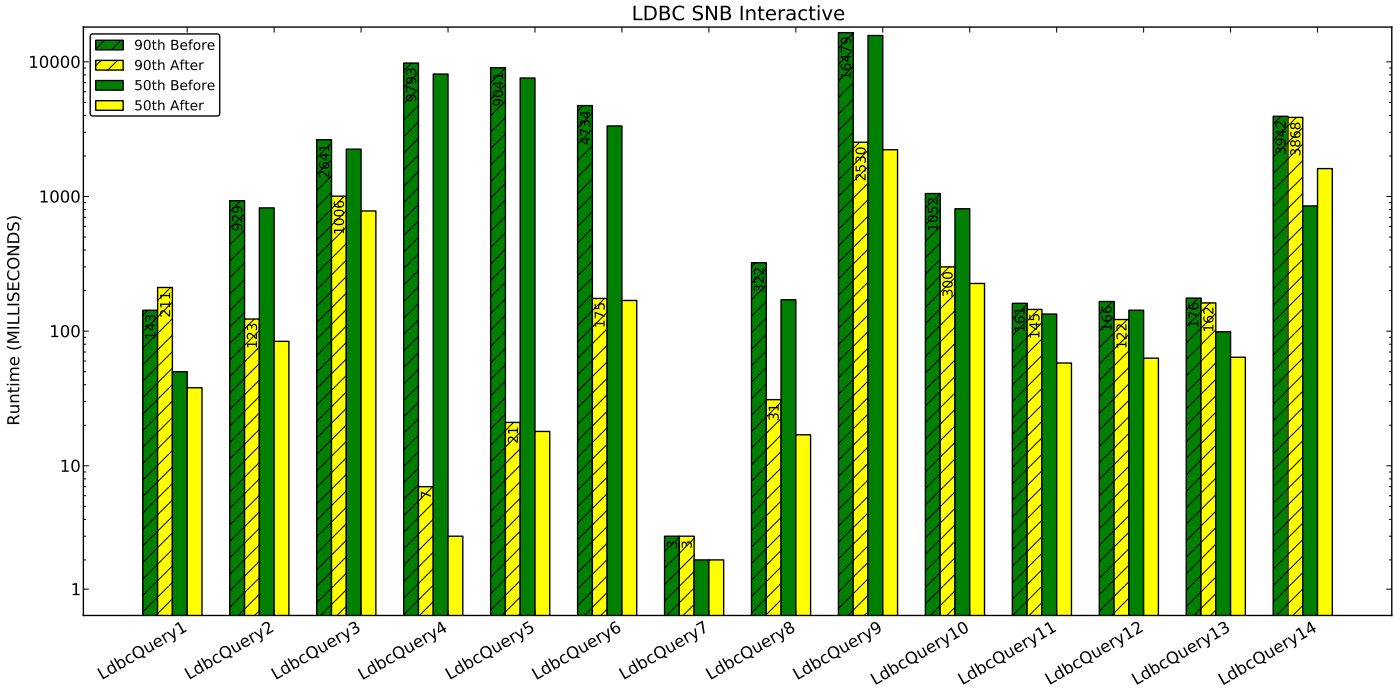
SNB in Sparksee - Things learned so far

- Exploiting correlations with Sparksee internal ids can be beneficial
  - Many queries ask for results sorted by date
  - Sparksee "Objects" collection is always sorted by id
  - We can avoid performing sorts
- Will these optimizations be valid for larger SFs? and updates?

SNB in Sparksee - Things learned so far

- Exploiting correlations with Sparksee internal ids can be beneficial
  - Many queries ask for results sorted by date
  - Sparksee "Objects" collection is always sorted by id
  - We can avoid performing sorts
- Will these optimizations be valid for larger SFs? and updates?
- The conclusion that we draw is that SNB interactive is rich, challenging and fun to play with

SNB in Sparksee - Things learned so far



Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates

Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates
- Add BI workload draft

Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates
- Add BI workload draft
- Improve on performance metrics

Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates
- Add BI workload draft
- Improve on performance metrics
- Add execution rules

Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates
- Add BI workload draft
- Improve on performance metrics
- Add execution rules
- Improve how-to's (running and validating)

Future of SNB - Specification

- Current version 0.2 $\rightarrow$ Just added interactive updates
- Add BI workload draft
- Improve on performance metrics
- Add execution rules
- Improve how-to's (running and validating)
- Please implement it and give us Feedback!

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability
- Hadoop 1.2.1 $\rightarrow$ 2.5.1

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability
- Hadoop 1.2.1 $\rightarrow$ 2.5.1
- Add flexibility
  - Custom serializers
  - Activate/Disable parts of the schema
  - Distribution overriding?

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability
- Hadoop 1.2.1 $\rightarrow$ 2.5.1
- Add flexibility
  - Custom serializers
  - Activate/Disable parts of the schema
  - Distribution overriding?
- Improve realism of the generated data

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability
- Hadoop 1.2.1 $\rightarrow$ 2.5.1
- Add flexibility
  - Custom serializers
  - Activate/Disable parts of the schema
  - Distribution overriding?
- Improve realism of the generated data
- Automatic statistics reporting

Future of SNB - DATAGEN

- Working on improving the codebase, documentation and performance/scalability
- Hadoop 1.2.1 $\rightarrow$ 2.5.1
- Add flexibility
  - Custom serializers
  - Activate/Disable parts of the schema
  - Distribution overriding?
- Improve realism of the generated data
- Automatic statistics reporting
- Correctness checking

Future of SNB - Driver

- Multiprocess support

Future of SNB - Driver

- Multiprocess support
- Built-in Warmup support

Future of SNB - Driver

- Multiprocess support
- Built-in Warmup support
- Short reads

Future of SNB - Driver

- Multiprocess support
- Built-in Warmup support
- Short reads
- Update validation process

Questions?

Thank you