Time Series Data Management, Distributed Applications, and Advanced Analytics

Presentation to Schlumberger

By:
Dr. Ramesh K. Raghunathan
ramesh.k.raghunathan@gmail.com
214.620.1863
September 25, 2015

Squigglee

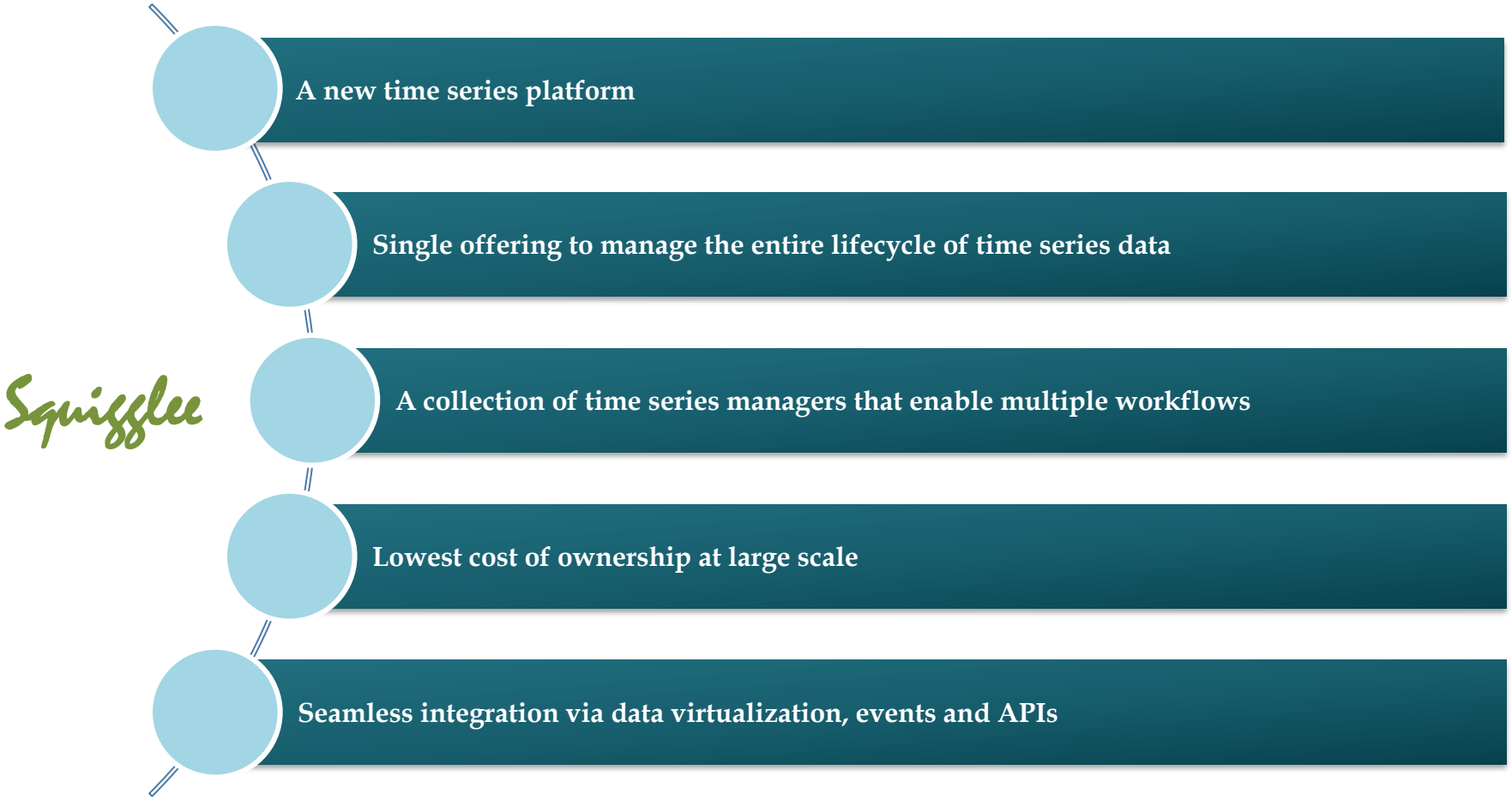
Presentation Agenda

...Time Series Done Right

- **Theme**
  - High frequency time series data management, distributed applications, and advanced analytics
- **Background**
  - This presentation is based on a product (Squigglee) that I built, and filed for patent, after leaving TCS
  - Product versions are currently in sales discussions
- **Outline**
  - Motivation & Challenges
  - Differentiators & Technology
  - Architecture & Capabilities
  - Demonstration Screenshots

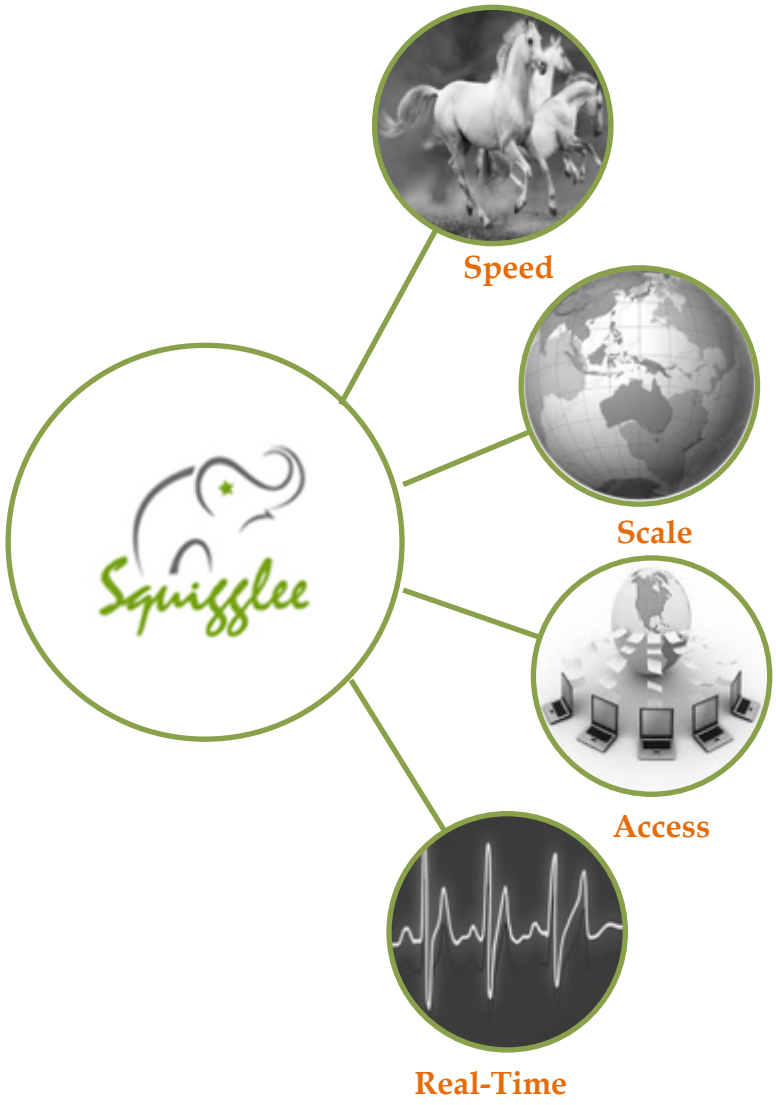
Summary

...Time Series Done Right



Platform Differentiators

...Time Series Done Right



- Multi-dimensional (pattern) indexes
- Sketch based data synopses
- On demand sampling
- Data locality – one time series entirely at one logical location

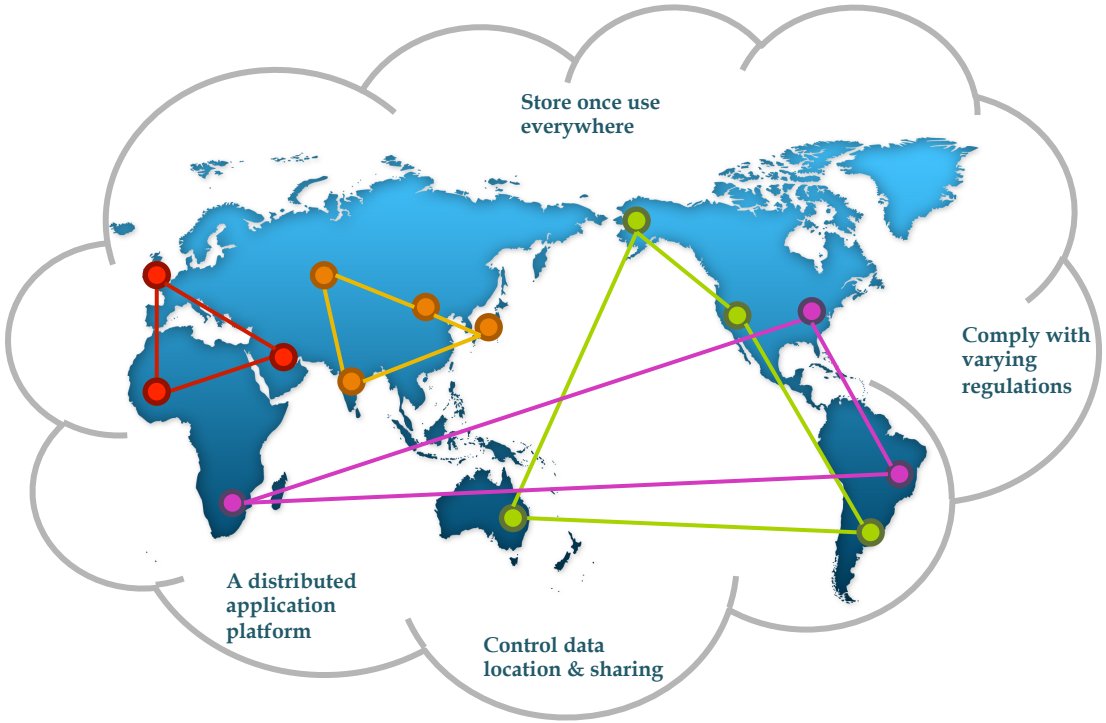
- Memory backed pre-allocated arrays in binary files
- Multi-frequency and high frequency support
- Node sets across heterogeneous data centers can be added or removed as desired

- Tiered distributed coordination w/ entitlement support
- Data virtualization layer
- REST proxies, query routing
- Data API w/ bulk upload and download

- Distributed complex event processing
- Global real-time replication -- variable by owner, data type, frequency, or need
- Incremental indexing & sketching

Sample Deployment Scenarios

...Time Series Done Right

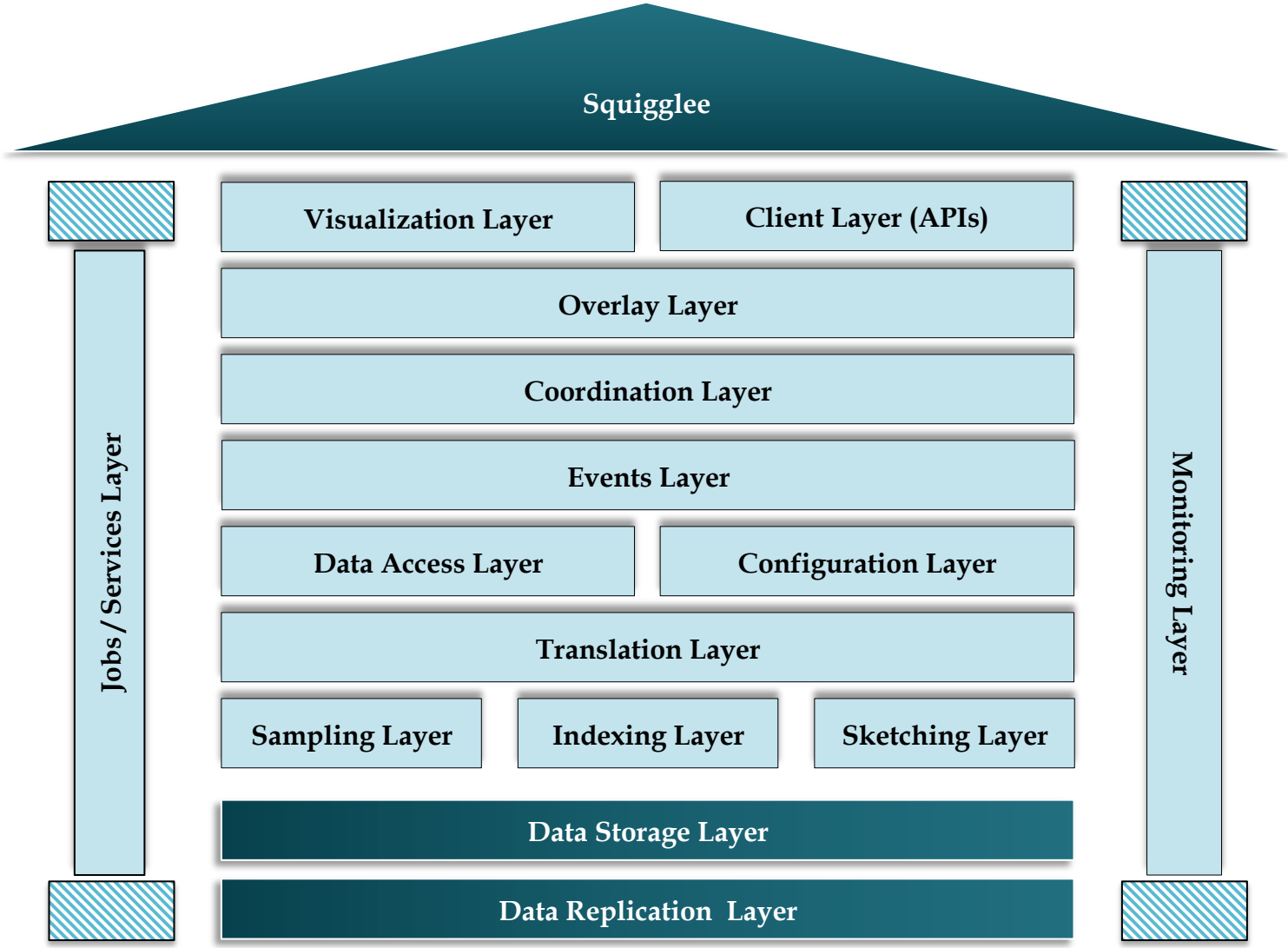


Sample Industry Applications

1. A conglomerate can ensure data replication occurs only to permitted or desired locations across countries (pursuant to local government regulations) while still enabling entitlement driven shared access.
2. An oil field services company accessing its own shared data for queries from multiple client locations, each of whom may have different data center providers (e.g. BP or Shell data centers located in Amazon or Google cloud locations respectively)
3. Governmental agencies sharing data for queries while still maintaining ownership and control over their data.
4. Medical industry institutions sharing data for specific research purposes without having to violate any patient or country health regulation.

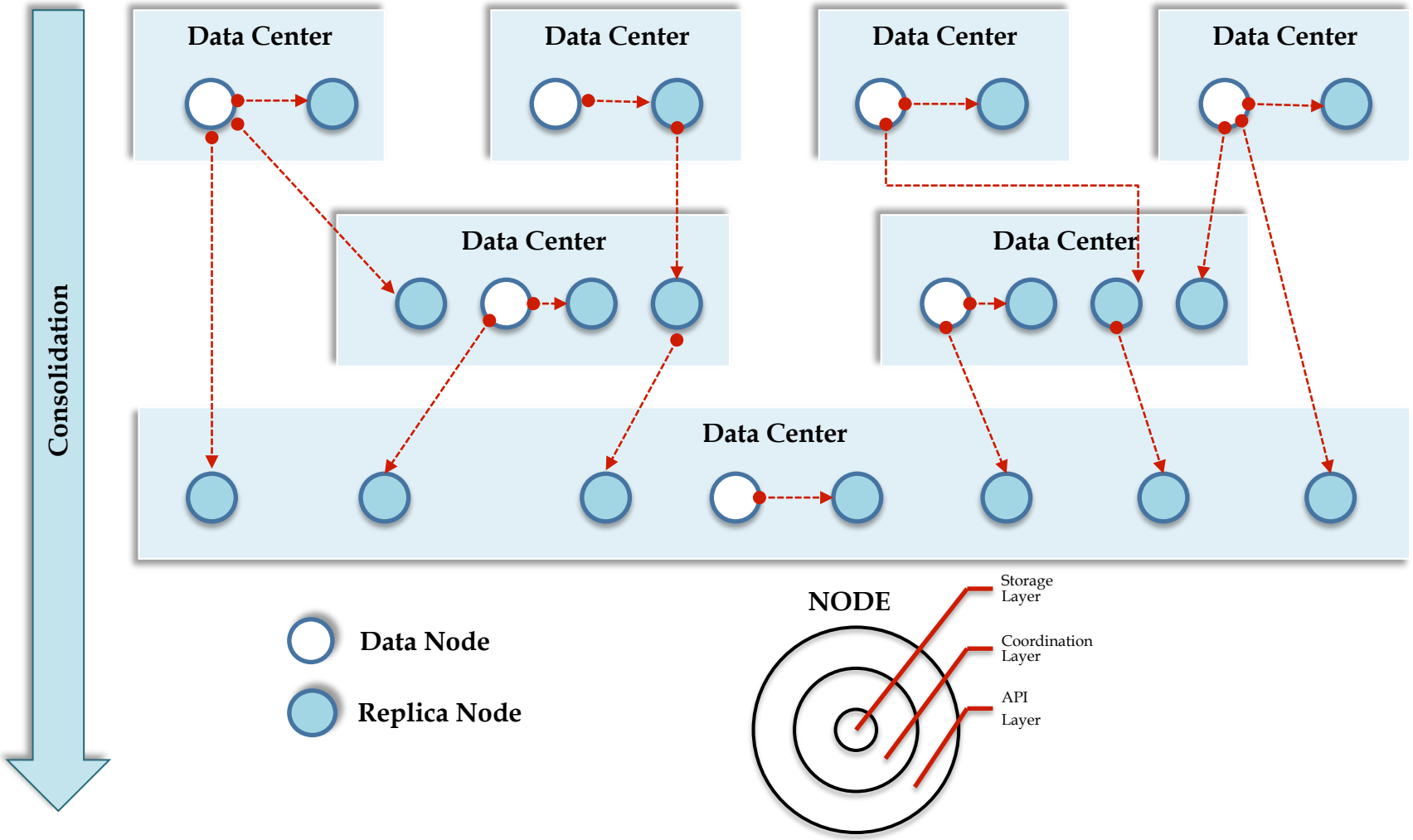
Summary Architecture

...Time Series Done Right



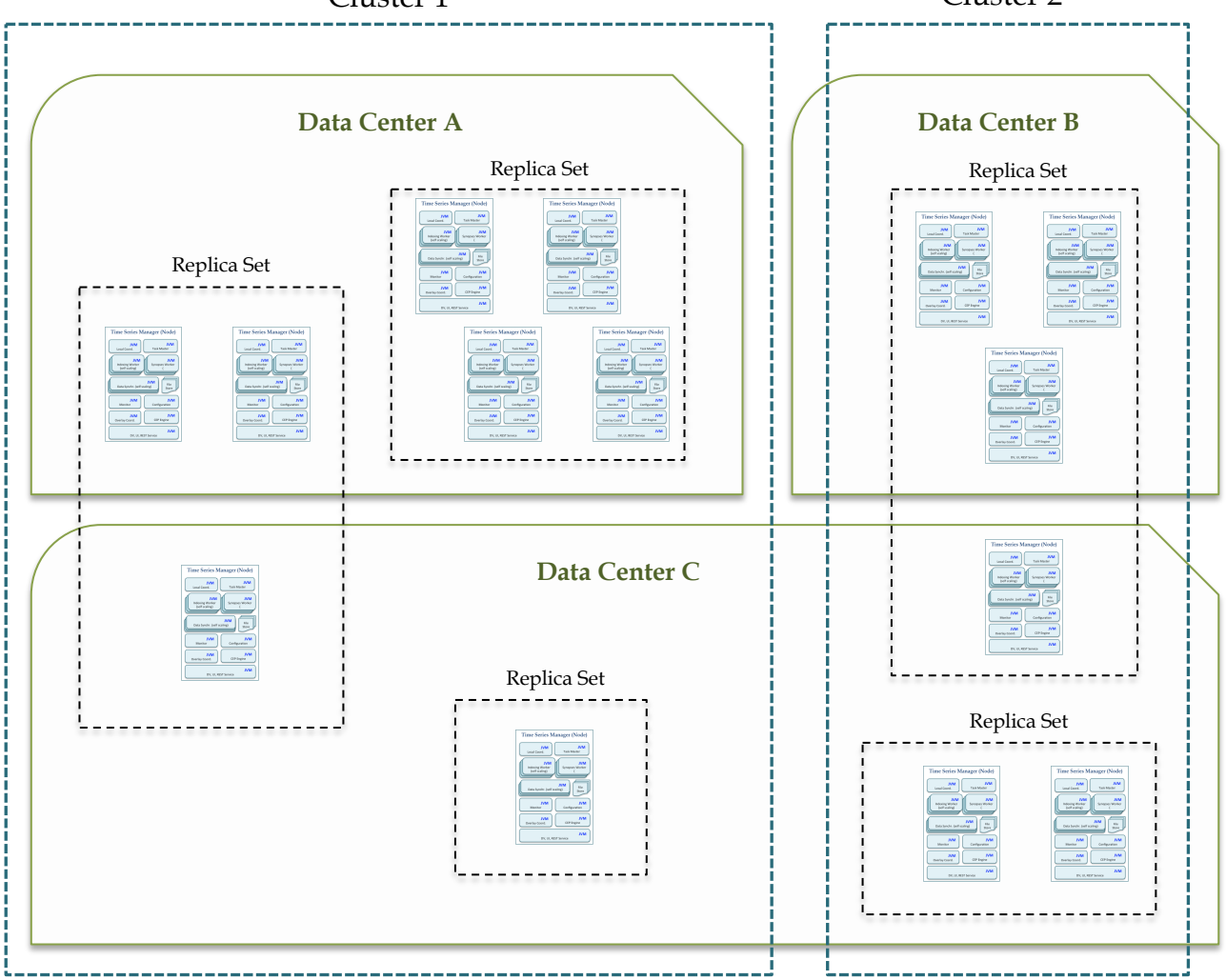
Deployment Illustration

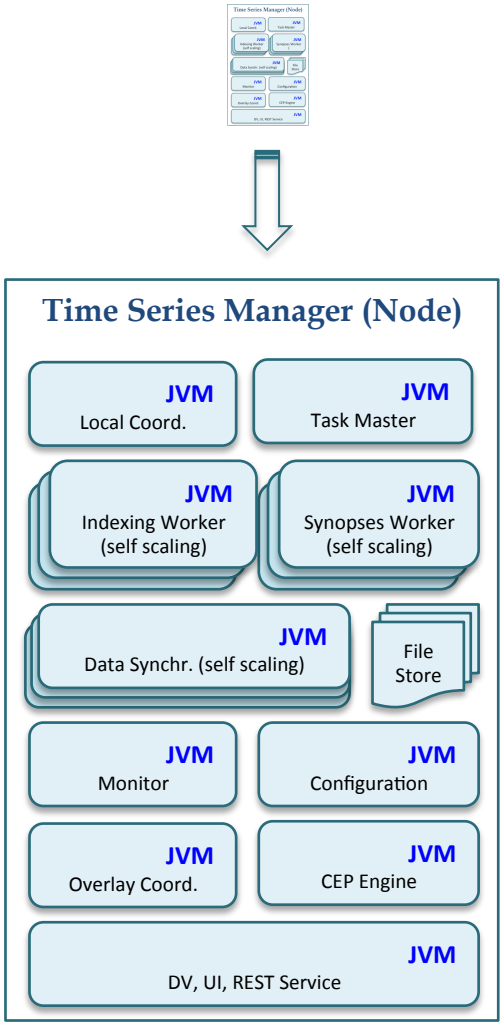
...Time Series Done Right



Deployed Artifacts

...Time Series Done Right





Pattern Matching

Squigglee

...Time Series Done Right

- **The problem**
  - Locate matches to any desired subsequence (motif) of a time series
  - Collections of motifs constitute domain competitive advantage
  - Avoid naïve brute force approaches approaches for searching large time series data sets
- **Challenges**
  - Must search matches in (near) real-time
  - Provide constant fast match retrieval regardless of size of search space
  - Must accommodate distributed data & indexes
  - Parameters – size of pattern (dimension), closeness of match, number of matches
  - Combine matches across multiple time series, save collections of patterns of interest, incrementally index in real-time
- **Solutions**
  - Dimensionality Reduction -- Easy to implement, scales poorly, too many matches, hierarchical piece-wise aggregation data structure
  - Locality Sensitive Hashing – Harder to implement, scales well, accurate matching but requires large storage, probabilistic data structure using a variety of hashing techniques

Data Synopses

- **The problem**
  - Provide exact & approximate answers quickly in real-time for large scale data
  - Queries of interest include basic statistics, frequency distributions, heavy hitters, count of distinct
- **Challenges**
  - Provide constant time retrieval regardless of size of search space
  - Incremental maintenance in (near) real-time
  - Point, range, & inverse queries must be supported
- **Solutions**
  - Sketches, Sampling, Histograms, Wavelets
  - Synopses data structures much smaller & roughly constant size regardless of the data size
  - Most common approaches use probabilistic counting data structures employing a variety of hashing techniques

Squigglee In Action

Squigglee

Home

...Time Series Done Right

Squigglee | Home | Operation | Configuration | Retrieval | Synopses

Operation

- Deploy globally
- Use replica sets to support any topology
- Control data lifecycle
- Share data with others

Configuration

- Add location as an attribute of time series
- Support multiple and high frequencies
- Support any data type

Retrieval

- Perform multi-dimensional data retrieval (pattern queries)
- Configure pattern indexes
- Capture and store patterns of interest
- Retrieve any number of matches within a radius

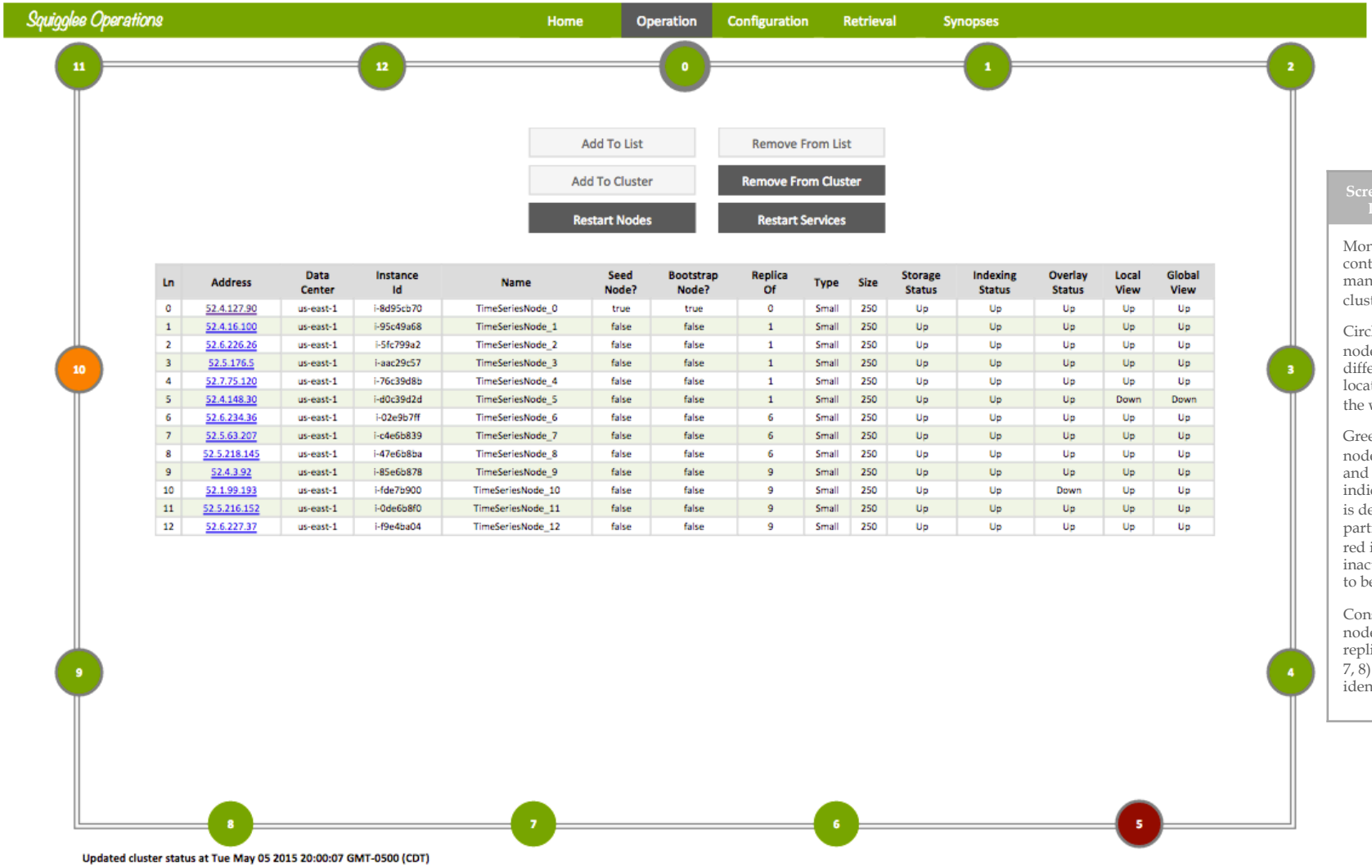
Synopses

- Configure sketches on large time series
- Query sub-samples on demand
- View summary statistics
- Perform point, range, and inverse estimates

Squigglee

Operation

...Time Series Done Right



Ln	Address	Data Center	Instance Id	Name	Seed Node?	Bootstrap Node?	Replica Of	Type	Size	Storage Status	Indexing Status	Overlay Status	Local View	Global View
0	52.4.127.90	us-east-1	i-8d95cb70	TimeSeriesNode_0	true	true	0	Small	250	Up	Up	Up	Up	Up
1	52.4.16.100	us-east-1	i-95c49a68	TimeSeriesNode_1	false	false	1	Small	250	Up	Up	Up	Up	Up
2	52.6.226.26	us-east-1	i-5fc799a2	TimeSeriesNode_2	false	false	1	Small	250	Up	Up	Up	Up	Up
3	52.5.176.5	us-east-1	i-aac29c57	TimeSeriesNode_3	false	false	1	Small	250	Up	Up	Up	Up	Up
4	52.7.75.120	us-east-1	i-76c39d8b	TimeSeriesNode_4	false	false	1	Small	250	Up	Up	Up	Up	Up
5	52.4.148.30	us-east-1	i-d0c39d2d	TimeSeriesNode_5	false	false	1	Small	250	Up	Up	Up	Down	Down
6	52.6.234.36	us-east-1	i-02e9b7ff	TimeSeriesNode_6	false	false	6	Small	250	Up	Up	Up	Up	Up
7	52.5.63.207	us-east-1	i-c4e6b839	TimeSeriesNode_7	false	false	6	Small	250	Up	Up	Up	Up	Up
8	52.5.218.145	us-east-1	i-47e6b8ba	TimeSeriesNode_8	false	false	6	Small	250	Up	Up	Up	Up	Up
9	52.4.3.92	us-east-1	i-85e6b878	TimeSeriesNode_9	false	false	9	Small	250	Up	Up	Up	Up	Up
10	52.1.99.193	us-east-1	i-fde7b900	TimeSeriesNode_10	false	false	9	Small	250	Up	Up	Down	Up	Up
11	52.5.216.152	us-east-1	i-0de6b8f0	TimeSeriesNode_11	false	false	9	Small	250	Up	Up	Up	Up	Up
12	52.6.227.37	us-east-1	i-f9e4ba04	TimeSeriesNode_12	false	false	9	Small	250	Up	Up	Up	Up	Up

Updated cluster status at Tue May 05 2015 20:00:07 GMT-0500 (CDT)

Screenshots from Live Demo

Monitoring and controls to manage a global cluster

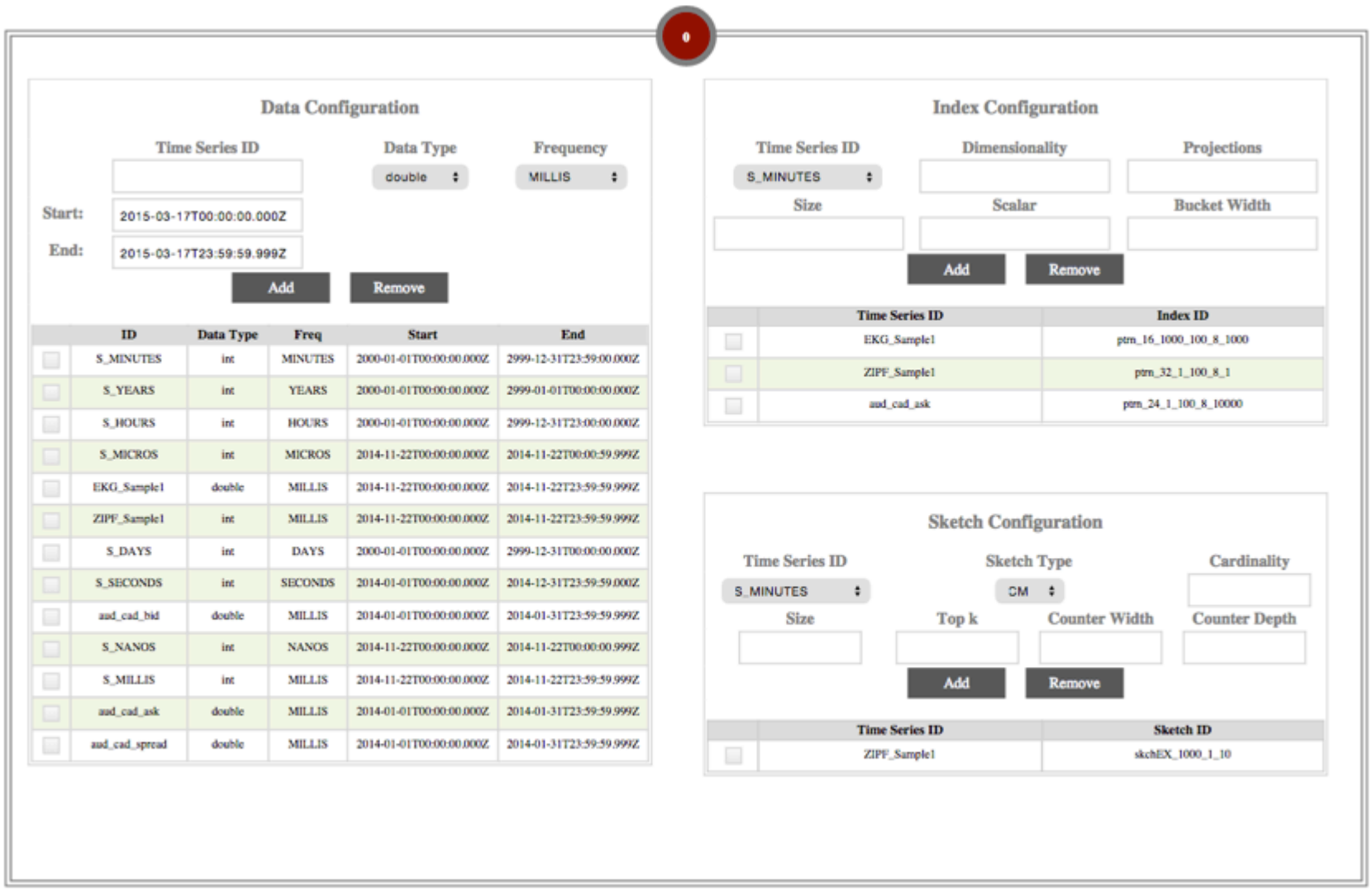
Circles represent nodes from different storage locations across the world

Green indicates – node is deployed and active, orange indicates -- node is deploying and partially active, red indicates inactive or ready to be deployed

Consecutive nodes represent replica sets (e.g. 6, 7, 8) each with identical data

Configuration

...Time Series Done Right



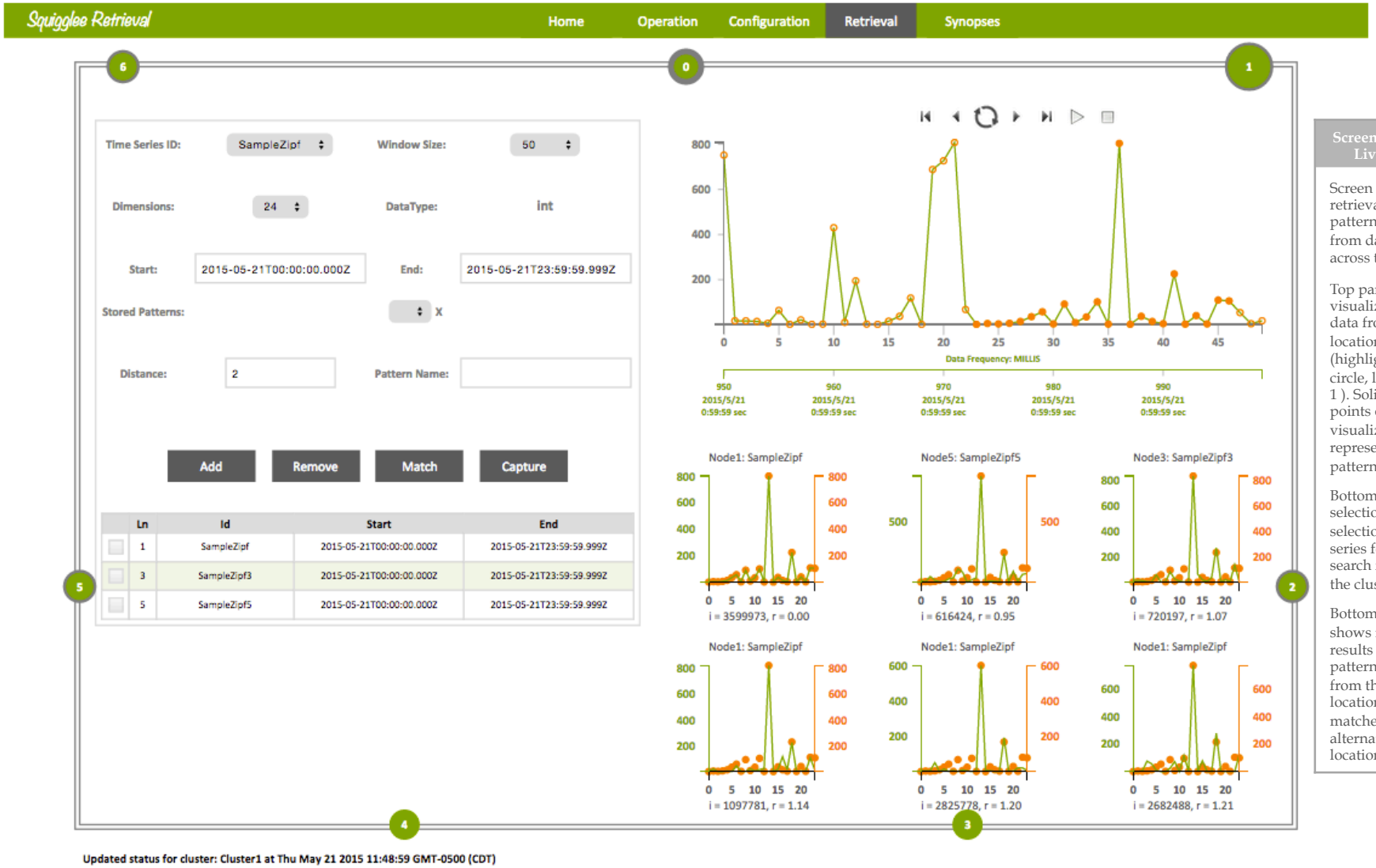
Updated cluster status at Mon Mar 16 2015 20:02:48 GMT-0500 (CDT)

Screenshots from Live Demo

- Screen allows configuration of nodes in the cluster
- Configure time series data storage intervals, frequency, and data types
- Configure pattern matching indexes
- Configure multiple sketches for data synopses

Retrieval

...Time Series Done Right



Screenshots from Live Demo

Screen allows data retrieval and pattern matching from data stored across the globe

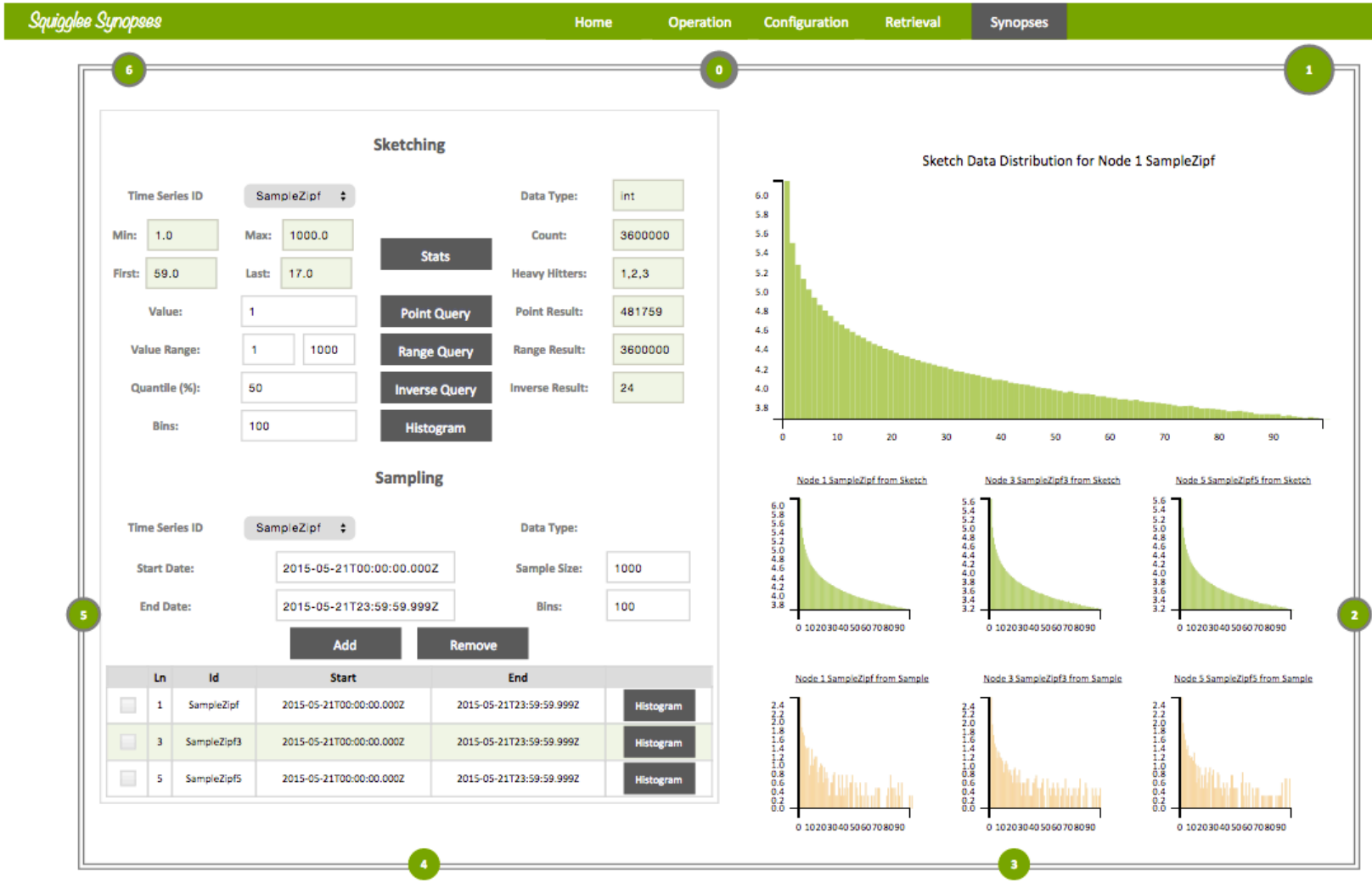
Top panel visualizes stored data from selected locations (highlighted circle, larger size, 1). Solid colored points on visualization represent a pattern of interest

Bottom left selection allows selection of time series for pattern search from any of the cluster nodes

Bottom right shows match results for the pattern of interest from the selected locations (e.g. matches from 3 alternate global locations shown)

Synopses

...Time Series Done Right



Screenshots from Live Demo

Screen show summary results of time series data as calculated from sketches and sampling

Top left panel shows summary statistics

Top right shows a frequency distribution of a specific time series from a selected node

Bottom left allows users to select time series data for sampling and analysis

Bottom right panel illustrates querying using random samples for fast real-time analyses

NOTES: Querying sketches & sub-sampled data much faster than processing all data

Thank You

For More Information Contact:
Dr. Ramesh K. Raghunathan
@ 214-620-1863
ramesh.k.raghunathan@gmail.com

Squigglee